ANÁLISIS EXPLORATORIO DE DATOS APLICADO A UNA BASE DE DATOS DE LOS SISMOS REGISTRADOS DEL AÑO 1970 AL AÑO 2022

EXPLORATORY DATA ANALYSIS APPLIED TO A DATABASE OF EARTHQUAKES RECORDED FROM 1970 TO 2022

Jonathan Leonel Cruz Ruiz
Tecnológico Nacional de México / IT de Celaya, México
m2303011@itcelaya.edu.mx

José Antonio Vázquez López
Tecnológico Nacional de México / IT de Celaya, México
antonio.vazquez@itcelaya.edu.mx

Edgar Augusto Ruelas Santoyo
Tecnológico Nacional de México / IT de Celaya, México
edgar.ruelas@itcelaya.edu.mx

Luis Gerardo Esparza Diaz
Tecnológico Nacional de México / IT de Celaya, México
gerardo.esparza@itcelaya.edu.mx

Recepción: 17/julio/2024 **Aceptación:** 24/febrero/2025

Resumen

Hoy en día la estadística es una herramienta importante en la vida académica como laboral. El estudio de la estadística con apoyos visuales es fundamental para poder entender los problemas que se presentan en la comunidad y en la vida diaria. En esta investigación abordo un fenómeno cuyo estudio es complejo, pero con la metodología a utilizar este fenómeno pudo ser estudiado y entendido con mayor facilidad, se trata del análisis exploratorio de datos. El análisis exploratorio de datos es una herramienta estadística que, con ayuda de gráficos, ayuda a analizar los datos sin necesidad de hacerlo analíticamente. Se uso gráficos de dispersión e histogramas para comprender una base de datos con datos históricos de 52 años y, por último, se describió lo observado a manera de poder comprender este fenómeno. El presente trabajo, dentro de los resultados y conclusiones obtenidos se observa la importancia y el impacto que el análisis exploratorio de datos tiene

sobre este tipo de fenómenos, el cual es muy importante en el campo de la investigación ya que es uno de los que más afecta a las sociedades.

Palabras Clave: Análisis exploratorio de datos, gráficas, sismos.

Abstract

Nowadays, statistics is an important tool in academic and professional life. The study of statistics with visual aids is essential to understand the problems that arise in the community and in daily life.

In this research, I address a phenomenon whose study is complex, but with the methodology to be used, this phenomenon could be studied and understood more easily, it is about exploratory data analysis. Exploratory data analysis is a statistical tool that, with the help of graphics, helps to analyze data without having to do it analytically. Scatter graphs and histograms were used to understand a database with historical data from 52 years and, finally, what was observed was described in order to understand this phenomenon. This work, within the results and conclusions obtained, the importance and impact that exploratory data analysis has on this type of phenomena is observed, which is very important in the field of research since it is one of those that most affects societies.

Keywords: *Exploratory data analysis, graphs, earthquakes.*

1. Introducción

En la actualidad la estadística es una herramienta útil para resolver diversos problemas en la vida diaria. En la presente investigación se aborda una técnica de estadística gráfica a un fenómeno natural de gran impacto como son los sismos, el esto con la finalidad de entender como ha sido el comportamiento a lo largo de los años. La investigación de sismos es importante ya que es uno de los fenómenos naturales más desastrosos que existen en la actualidad y poder entender su comportamiento abre las puertas a investigaciones que apoyen a la prevención.

[Batanero, Estepa & Godino, 1991] comentan que en la actualidad la enseñanza de la Estadística se realiza de forma gradual desde la educación básica hasta alcanzar niveles de licenciatura, también comentan que debido al espectacular desarrollo de

la informática y a la disponibilidad de paquetes de cálculo, fácilmente manejables y accesibles, asistimos en nuestros días a una demanda cada vez mayor de formación estadística. Por otro lado, [Tukey, 1977] dice que las capacidades de cálculo y representación gráfica de los ordenadores actuales permiten de una forma sencilla, la obtención de una amplia variedad de gráficos y estadísticos diferentes y han hecho posible la aparición de una nueva filosofía en los estudios estadísticos: el análisis exploratorio de datos.

El Análisis Exploratorio de Datos (EDA por sus siglas en inglés) es una fase crítica en el proceso de análisis de datos que permite a los analistas explorar y comprender mejor la estructura y las características de los datos. Introducido por John Tukey, el EDA combina técnicas estadísticas y visuales para identificar patrones, detectar anomalías, probar hipótesis preliminares y resumir las principales características de los datos sin realizar suposiciones previas. Las técnicas del EDA incluyen:

- Visualización de Datos: Uso de gráficos como histogramas, diagramas de caja, gráficos de dispersión, gráficos de violín y gráficos de líneas para identificar patrones y relaciones entre variables.
- Estadísticas Descriptivas: Cálculo de medidas como la media, la mediana, la desviación estándar, el rango y los percentiles para resumir la distribución y la variabilidad de los datos.
- Detección de Valores Atípicos: Identificación de puntos de datos que se desvían significativamente del resto del conjunto de datos, lo cual puede indicar errores o fenómenos interesantes.
- Análisis Multivariante: Exploración de relaciones entre múltiples variables mediante el uso de matrices de correlación, gráficos de dispersión matriciales y análisis de componentes principales.

También es importante explicar que el EDA es esencial para:

- Comprender la Distribución de Datos: Identificar la forma y los patrones de distribución para comprender mejor los datos.
- Verificar Suposiciones: Evaluar si los datos cumplen con las suposiciones necesarias para análisis estadísticos más avanzados.

- Preparación de Datos para Modelización: Detectar y manejar valores atípicos y datos faltantes, y realizar transformaciones necesarias para mejorar la interpretación y el rendimiento de los modelos.

El EDA es un paso preliminar pero fundamental en el análisis de datos que facilita una comprensión profunda de los datos y guía las decisiones analíticas y de modelización subsiguientes [Galit, Nitin & Peter, 2010].

De acuerdo con lo propuesto por [Hoaglin, Mosteller, & Tukey, 1983] y por [Velleman y Hoaglin, 1981] se reconoce la existencia de cinco características principales del EDA:

- Sus representaciones graficas nos revelan, en una primera fase, el comportamiento de los datos y la estructura del conjunto.
- Dedica mucha atención al análisis de residuales.
- Utiliza la transformación de los datos para conseguir ajustar los valores originales a la escala que más simplifique y clarifique el análisis como, por ejemplo, mediante el uso de funciones matemáticas simples (raíz cuadrada, logaritmos, etc.).
- Valora la resistencia, propiedad que presentan algunos estadísticos que les hace poco sensibles a la influencia de uno o varios valores marcadamente distantes de la mayoría de los datos de la distribución.
- Busca estadísticos robustos, propiedad que presentan algunos estadísticos que les hace poco sensibles ante desviaciones de los supuestos básicos.

Para entrar más en contexto, [Filiben & Heckert, 2012] mencionan que el análisis de datos exploratorios (EDA) es un enfoque/filosofía para el análisis de datos que emplea una variedad de técnicas (principalmente gráficas) para

- Maximizar el conocimiento de un conjunto de datos.
- Descubrir la estructura subyacente.
- Extraer variables importantes.
- Detectar valores atípicos y anomalías.
- Probar los supuestos subyacentes.

- Desarrollar modelos parsimoniosos.
- Determinar la conFiguración óptima de los factores.

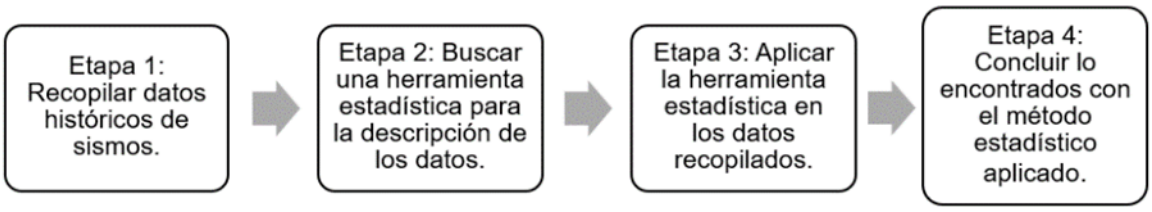
Los autores también comentan los gráficos que se pueden utilizar para el análisis exploratorio de datos. Algunos ejemplos son: histogramas, gráficos de caja y bigotes, gráficos de probabilidad, gráficos lineales, gráficos de autocorrelación, gráficos de dispersión, entre otros. El tipo de grafico a usar depende del problema a bordar en la investigación.

Otro factor importante por estacar es que hay un paso diferente a la estadística clásica: en la estadística clásica primero se hace el modelo y después se analiza y en el análisis exploratorio de datos primero se analizan los datos y luego se crea el modelo.

2. Método

Dentro de esta investigación se aborda el análisis exploratorio de datos enfocado a el fenómeno de los sismos, esto con la finalidad de entender el comportamiento de estos fenómenos en los años analizados en la base de datos.

La Figura 1 muestra un diagrama con las 4 etapas que se siguieron para este trabajo, metodología a seguir para llevar a cabo la investigación.



Fuente: elaboración propia.

Figura 1 Diagrama de las etapas llevadas a cabo en este trabajo.

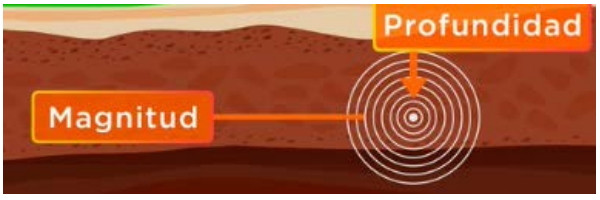
En la etapa 1 que es la recopilación de datos, se trabajó con una base de datos históricos de 52 años (1970 a 2022) recopilados de la plataforma Seismic monitor creada por el consorcio IRIS (Incorporated research institutions for sismology).

Para la etapa 2 se buscó diferentes opciones estadísticas para tratar los datos y se optó por utilizar el análisis exploratorio de datos. Se trabajó solo con dos variables

de la base de datos: la magnitud (que es la fuerza liberada por el sismo) y la profundidad (que es la distancia que se ve afectada por un sismo), esto debido a que estas dos variables son las que, en conjunto, pueden provocar grandes desastres dentro de las comunidades. Para esta etapa se utilizó el software "Minitab", en este software estadístico se introdujo toda la información de la base de datos y una vez ingresada fue procesada.

La Figura 2 muestra un diagrama de cómo se visualiza la magnitud y profundidad sísmica. Se observa la magnitud y la profundidad de un sismo, la magnitud se visualiza como la energía liberada por un sismo y se observa a manera de ondas la profundidad es que tan adentro de la corteza terrestre comienza el sismo.

El análisis exploratorio de datos consta de diferentes visualizaciones estadísticas para el análisis de los datos, pero para llevar a cabo la etapa 3 se trabajó particularmente en gráficos de dispersión e histogramas y a pesar de que la nace de datos cuenta con 52 años, solo se presentan los resultados de dos años extremos (1970 y 2022) y un dato intermedio (2000), esto para no hacer más digerible el trabajo y no atiborrarlo de cada uno de los años que tiene la base de datos. Por último, para la etapa 4 se revisaron los diferentes gráficos y se tomaron consideraciones de lo observado en cada uno de ellos. Esta última etapa seria lo que se conoce como conclusiones y están descritas al final del documento en la sección 4. En la Figura 3 se observa un ejemplo de un gráfico de dispersión mientras que en la Figura 4 se observa un ejemplo de un histograma.

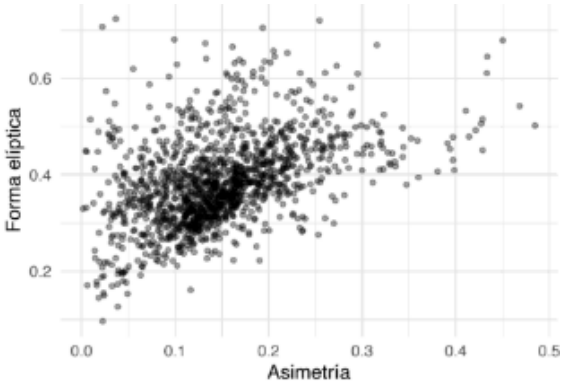


Fuente: SkyAlert.

Figura 2 Diagrama magnitud y profundidad.

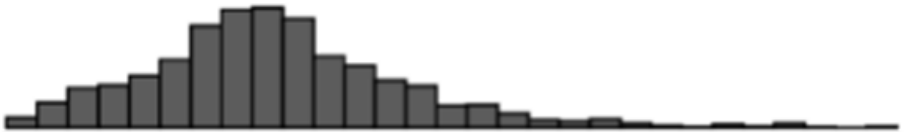
3. Resultados

Primero se trabajó con gráficos de dispersión de los años previamente mencionados (1970, 2000 y 2022).



Fuente: Herramientas para el análisis estadístico de datos biológicos en R.

Figura 3 Ejemplo de grafico de dispersión.



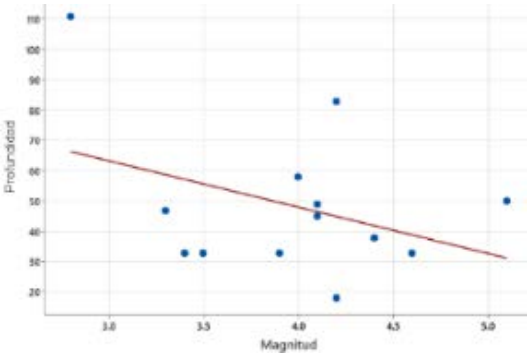
Fuente: Herramientas para el análisis estadístico de datos biológicos en R.

Figura 4 Ejemplo de histograma.

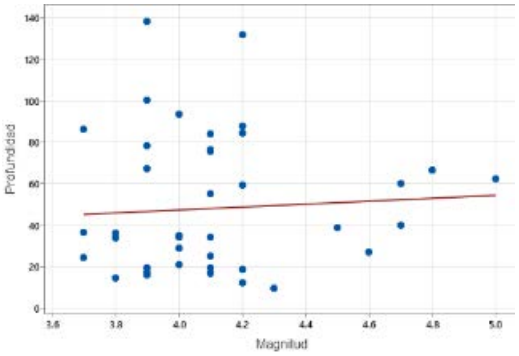
En la Figura 5 se muestran las gráficas de dispersión del año, en la Figura 5a del año 1970, se observa que los valores de la magnitud comienzan en 2.5 y terminan en 5.5 (aproximadamente), no se observa ninguna relación entre ambos parámetros. En la Figura 5b del año 2000, se observa que las magnitudes van de 3.6 a 5.0 aproximadamente, no se observa ninguna relación entre ambos parámetros. En la Figura 5c del año 2022, se observa que en las magnitudes por debajo de 4 se mantienen en niveles bajos de profundidad, pero de una magnitud mayor a 4 el nivel de profundidad aumenta.

En la Figura 6 se muestran los histogramas de magnitud varios años. En la Figura 6a del año 1970 se observa que no se encuentra sesgada y el mayor número de eventos estuvieron en valores “bajos” de la magnitud. Se observa que tiende a un comportamiento normal en la magnitud de 4. En la Figura 6b del año 2000 se observa que se encuentra sesgada a la izquierda, esto indica que la mayoría de los eventos se llevaron a cabo en valores “bajos” de la magnitud. En la Figura 6c del año 2022, se observa que no se encuentra sesgada y que muestra un comportamiento de subida y bajada dentro de los valores “bajos” de la magnitud. En

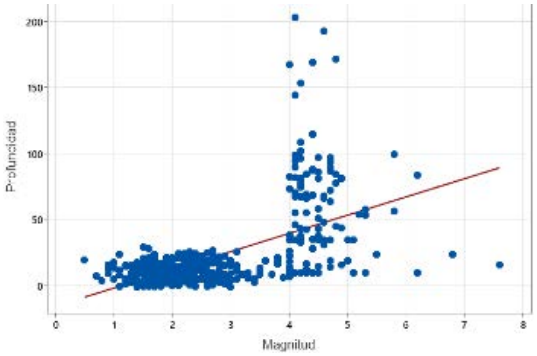
los histogramas de la Figura 6, se observa que independientemente del comportamiento de los eventos, la mayoría son de magnitudes “bajas”.



a) 1970.



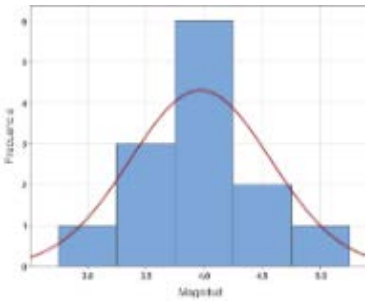
b) 2000.



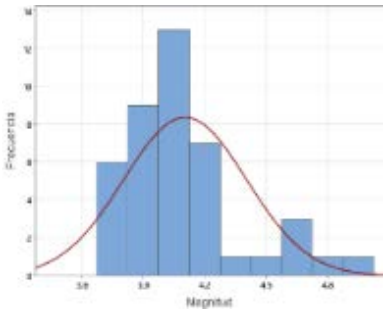
c) 2022.

Fuente: elaboración propia.

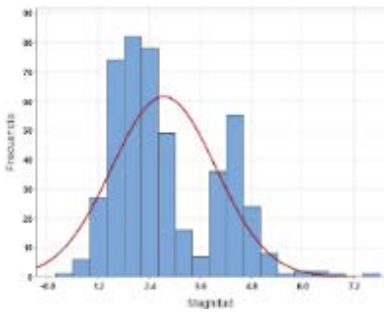
Figura 5 Gráficos de dispersión Profundidad vs Magnitud en varios años.



a) 1970.



b) 2000.



c) 2022.

Fuente: elaboración propia.

Figura 6 Histogramas en cada año.

4. Discusión

Los resultados obtenidos en la aplicación del análisis exploratorio de datos a una base de datos sísmicos fueron satisfactorios ya que el análisis exploratorio de datos

arroja un panorama más digerible para el análisis de datos ya que permite observar anomalías entre los datos, si siguen alguna tendencia, su comportamiento, etc.
Si bien, solo se estudiaron a profundad tres años de la base de datos, en un trabajo posterior se puede trabajar más a detalle con todos los años para así observar el comportamiento y las anomalías presentes en cada uno de los años.

5. Conclusiones

El análisis exploratorio de datos es una herramienta eficaz que ayuda al analista a describir mejor el comportamiento de los datos y arroja una visión de cómo van evolucionando a través del tiempo.
Para este trabajo en específico solo se trabajó con dos tipos de gráficos: gráficos de dispersión e histogramas, sin embargo, el análisis exploratorio de datos tiene muchas herramientas visuales.
Para el caso de los gráficos de dispersión, los años más lejanos a la actualidad (1970 y 2000) no presentan muchos datos y por eso no se puede observar claramente el comportamiento que tienen, sin embargo, los datos que existen muestran que no existe ningún tipo de relación entre los datos; por otro lado, para el año 2022 se observa que la magnitud 4 es crucial para el nivel de profundidad ya que en ese punto existe un aumento considerable en el gráfico, esto es un área de oportunidad que nos arroja este grafico para un trabajo posterior donde se podría indagar más acerca de esta anomalía en los datos. Respecto al otro gráfico utilizado que fue el histograma, se observa que en el año 1970 el gráfico sigue una tendencia normal a diferencia de los otros años estudiados, para el año 2000 se observa el grafico sesgado a la izquierda y para el año 2022 se observa que el histograma se empieza a dividir y se muestran dos histogramas diferentes.
Los resultados fueron satisfactorios ya que se pudo observar cómo es el comportamiento de los sismos a lo largo de los años, dentro de las herramientas de la estadística, el análisis exploratorio de datos ayuda al procesamiento de la información y ayuda a visualizar como se están comportando los datos para en investigaciones posteriores poder realizar metodologías para investigaciones de este tipo de fenómenos.

6. Bibliografía

[1] Batanero, C., Estepa , A., & Godino , J. D. (1991). Análisis exploratorio de datos: sus posibilidades en la enseñanza secundaria. Suma, 25-31

[2] Filliben , J., & Heckert, A. (2012). Exploratory Data Analysis . Engineering Statistics Handbook.

[3] Galit , S., Nitin , R. P., & Peter , C. B. (2010). Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. New Jersey: John Wiley & Sons.

[4] Hoaglin, D. C., Mosteller , F., & Tukey, J. W. (1983). Understanding Robust and Exploratory Data Analysis. New York: John Wiley & Sons.

[5] Tukey, J. W. (1977). Exploratory Data Analysis.

[6] Velleman, P. F., & Hoaglin, D. C. (1981). Applications, Basics and Computing of Exploratory Data Anaiysis. Boston: Duxbury Press.