MINERÍA DE DATOS APLICADA A SISMOS EN MÉXICO

MINING OF DATA APPLIED TO SISMOS IN MEXICO

Andrés Aristóteles Suárez Rojas[1] Doricela Gutiérrez Cruz[2] Ricardo Rico Molina[3] Angélica Caballero Hernández[4]

RESUMEN

La minería de datos es una técnica que consiste en la aplicación de algoritmos específicos que generan una enumeración de patrones a partir de grandes volúmenes de información, útiles para la toma de decisiones en amplios campos de aplicación. En este trabajo fueron analizados datos históricos sobre sismos registrados en los tres estados con mayor cantidad de registros en México por el Servicio Sismológico Nacional, utilizando series temporales, con el objeto de entender su comportamiento.

Palabras clave Minería de datos, Series temporales, coeficiente de Hurst.

ABSTRACT

Data mining is a technique that involves the application of specific algorithms that generate a list of patterns from large volumes of information, useful for decision-making in wider fields of application. In this paper we were analyzed historical data on earthquakes recorded in the three states with the highest number of registrations in Mexico by the National Seismological Service, using time series in order to understand their behavior.

Keywords: Data mining, time series, Hurst coefficient.

Unidad Académica Profesional Nezahualcóyotl-UAEM.

Recibido: 31-octubre-2016 / Aceptado: 15-diciembre-2016.

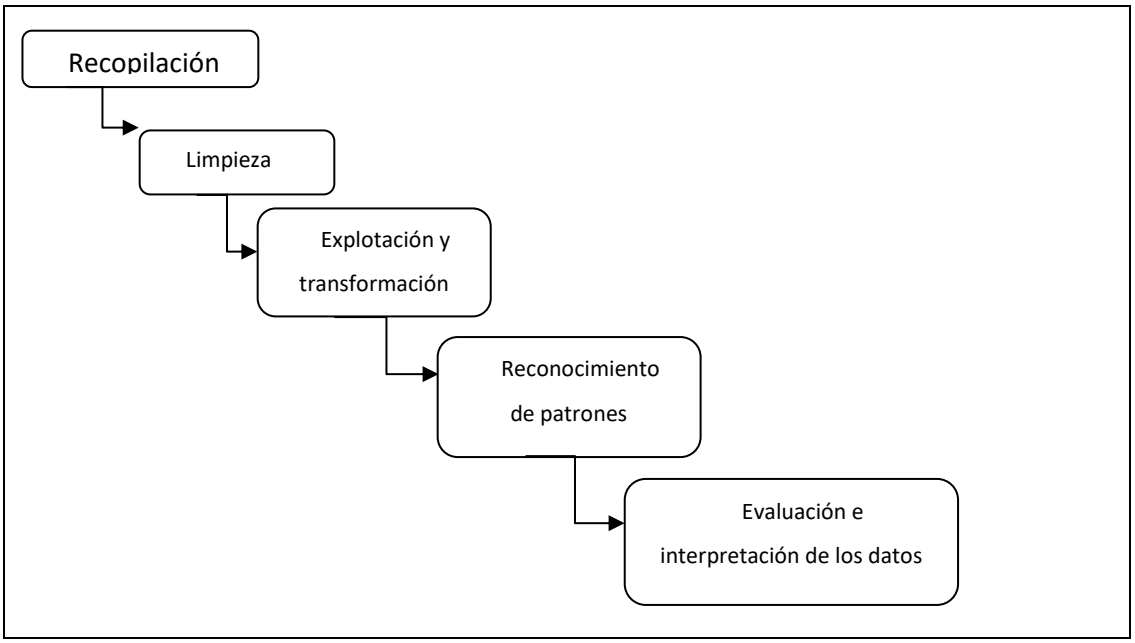
INTRODUCCIÓN

El volumen y diversidad de información que se encuentra computada en bases de datos digitales ha crecido considerablemente en la última década (Hernández, Ramírez y Ferri, 2004), parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido y puede ser de utilidad para comprender la información futura (Simon, 1997; Berson y Smith, 1997; White, 2001). Como es el caso de la minería de datos (MD) que se puede ver como un proceso en el cual se aplican algoritmos en específico para analizar fenómenos no visibles en grandes cantidades de datos, para encontrar patrones ocultos, relaciones entre variables y con esto obtener una descripción del comportamiento de los datos analizados (Zúñiga, 2016; Velarde, 2003; Aguirre, *et. al.,* 2015; Riquelme, *et. al*., 2006; Timarán, Calderón y Jiménez, 2013), a través del uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos pre-procesados, que sean de utilidad para la toma de decisiones en distintas áreas (Hand, Mannila y Smyth, 2001; Frawley, Piatetsky-Shapiro y Matheus, 1992; Fayyad, Piatetsky-Shapiro, *et. al.*, 1996), este proceso consta de cinco fases que se muestran en la Figura 1.

Los movimientos sísmicos se pueden explicar con la teoría de tectónica de placas, la cual sustenta que la superficie de la tierra está formada por placas rígidas que flotan

sobre un manto plástico y caliente que su longitud puede ser de miles de kilómetros y.

Figura 1.
Proceso de la minería de datos



Fuente: Elaboración propia, 2016.

de grosor es alrededor de 100 km, y estas placas se mueven 5 cm por año, al rosarse o chocar estas placas generan estos movimientos sísmicos y a largo plazo pueden llegar a generar grandes grietas en la tierra como la falla de san Andrés, la cordillera de los andes (Espinoza y Jiménez, 2016), concretamente la república mexicana está situada geográficamente en una de las regiones sísmicamente más activas del mundo según el Servicio Sismológico Nacional (SNN) (Gerardo y Zenón, 2016).

Parte de la aleatoriedad de los fenómenos naturales se debe a causas internas del sistema o a factores externos estocásticos, particularmente si estos no varían en forma lineal. El estudio de las variables y las interacciones de un sistema dinámico a través del tiempo se enfoca a encontrar patrones, estructuras y puntos críticos de estabilidad o inestabilidad así como la sensibilidad al cambio de las condiciones iniciales para lograr cierto grado de control (Balankin, Morales y Gálvez, 2004).

MARCO TEÓRICO

Un algoritmo que se utiliza en la MD es el de series temporales, con el cual se puede pronosticar o estimar el valor de un dato inmediato de una serie de datos que se está

analizando, para llevar a cabo este algoritmo se necesita contar con antecedentes históricos, los cuales se caracterizan por que están dispuestos a su tratamiento o proceso por computadora (Rodríguez, 2008; Botero y Cano, 2008). La administración de series de tiempo se ha convertido en un área de investigación importante en MD (Cáceres y Rodríguez, 2011) debido a que integra las técnicas de MD en un herramienta computacional para demostrar por métodos experimentales que funcionan correctamente (Rodríguez, 2006), la dinámica de las series de tiempo puede tener un comportamiento complejo, similar a un proceso estocástico, que bien se pueden analizar desde la geometría fractal, la cual estudia los aspectos geométricos que son invariantes con el cambio de escala, dentro de las diversas medidas no lineales para estimar la complejidad de una serie de tiempo se cuenta con una herramienta para evaluar dos atributos de gran relevancia en el estudio de la geometría fractal, como son el exponente de Hurst y la dimensión fractal Mandelbrot (Mandelbrot, 1988). Estos atributos se relacionan con el grado de rugosidad que puede llegar a presentar las series de tiempo. La dimensión fractal es una magnitud estadística que permite describir matemáticamente los objetos que presentan alto grado de complejidad, de auto-similaridad o caóticos. Con la estimación del exponente de Hurst y la dimensión fractal de las series de tiempo se puede analizar si una serie de tiempo es fractal y se puede comprobar si ésta tiene memoria.

La estimación del exponente de Hurst (H) se ha aplicado en áreas que van desde la biofísica a las redes de computadoras. El método del exponente de Hurst fue desarrollado originalmente para estudios hidrológicos, sin embargo las modernas técnicas para estimar el exponente de Hurst provienen de la matemática fractal. Esta estimación proporciona una medida para comprender si los datos son un camino aleatorio puro o tienen tendencias subyacentes, determina el grado de la aleatoriedad, el cual representa la persistencia de un fenómeno estadístico (Salas, Delleur, Yevjevich y Lane, 1985; Schroeder, 1991). En el caso de un fenómeno con comportamiento aleatorio puro, el coeficiente de Hurst tiene valor igual a 0.5; es decir, similar a la distribución Gaussiana o al movimiento Browniano clásico (Mandelbrot, 2002). El coeficiente de Hurst es un indicador de la rugosidad de la base de datos y los valores menores de 0.5, indican una tendencia de regresar en sí mismos, propiedad que es conocida como anti-persistencia y los valores mayores de 0.5, indican la tendencia a persistir en su progresión en la dirección que se está moviendo y se conoce como persistencia.

En Kagan, Jackson, Rong (2007), utilizaron un método, que está basado en un catálogo de espacialidad histórica de terremotos, para presentar una predicción a cinco años de terremotos de magnitud 5.0 o más, para el sur de California y su principal característica recae en la observación de regularidades en aparición de terremotos. Como se menciona

en Morales, Martínez, Troncoso, De Justo, Rubio (2010), utilizaron un algoritmo conocido como *clustering K-means*, para asociar algunos patrones con las variaciones del valor b. Los autores evalúan sus hipótesis sobre datos de la península Ibérica. Estos patrones son capaces de predecir a un medio plazo la aparición de terremotos con gran confiabilidad, por otra parte Martínez, Troncoso, Morales y Riquelme (2011), pero esta vez, usando el algoritmo M5P y reglas de asociación cuantificadas. Los autores mostraron la fuerte relación existente entre las variaciones negativas del valor "b" y los grandes terremotos.

Debido a las múltiples pruebas descubiertas, se decidió a que las variaciones del valor "b" fueran los datos de entrada de diferentes algoritmos aplicados. En referencia al uso de modelos neuronales, Adeli y Panakkat (2009), usan una red neuronal probabilística, este tipo de red neuronal es principalmente usada para clasificación de problemas, como se ha aplicado en este trabajo, en el cual se usan los datos de la región sur de California como datos de entrada, dando como resultado una predicción de la magnitud de los terremotos, como uno de los valores de salida de las clases.

MÉTODO

El estudio se realizó con información histórica que ofrece el servicio sismológico nacional (SNN, http://www.ssn.unam.mx/), acerca de los sismos que ocurren en la república mexicana mediante aparatos que miden la magnitud, profundidad y epicentro de los sismos, estos dispositivos se encuentran distribuidos en puntos específicos del país. En este caso se estudiará la magnitud de los sismos entre los años 2006 a 2015, en tres de los estados con mayor número de sismos registrados según el SSN; los cuales son: Chiapas, Guerrero y Oaxaca.

Recopilación y limpieza

Se utilizaron las series históricas de magnitud sísmica entre enero de 2006 a diciembre 2015 del SSN. Se elaboró una base de datos a nivel diario en Excel, a partir de la cual se generaron archivos para cada mes (escala mensual) y para cada año (escala anual).

Estos mismos archivos se guardaron como series de tiempo, para calcular la dimensión fractal y el coeficiente de Hurst, utilizando el método de referencia del rango re-escalado, diseñado para el análisis de los patrones auto-afines con el programa Benoit®.

Explotación y transformación

Teniendo el formato adecuado para trabajar con los datos, se generan gráficas para cada estado de los tres estados para ver qué tan volátiles son estas.

De los tres estados estudiados se puede apreciar que hay periodos y muchos puntos en los cuales la magnitud desciende hasta cero (cuando no se registran sismos), esto implica que no es tan sencillo comprender su comportamiento debido a la inestabilidad del valor de la magnitud de los sismos.

Reconocimiento de patrones

El coeficiente de Hurst determina la intensidad de la dependencia entre los datos y de acuerdo con su magnitud, la serie de tiempo se clasifica como persistente (0.5<H<=1), que significa que existe dependencia entre un evento y los ocurridos anteriormente; cuando se clasifica como anti persistente (0<=H<0.5) significa que en la serie persiste una tendencia a ser caótica o que sus valores tienen una alta volatilidad. En caso de que H = 0.5 se concluye que la serie de tiempo es aleatoria y los datos no se correlacionan entre sí; es decir, los valores futuros de la serie no son influenciados entre ellos por lo que ocurre el presente (Palomas, 2002).

Este último caso modela el ruido blanco, la distorsión Gaussiana normal o movimiento Browniano clásico. Los casos anteriores describen los movimientos Brownianos fraccionarios. El valor de H permite determinar si el comportamiento de datos de la precipitación es persistente o anti-persistente (Burgos y Pérez, 1999; Miranda, Andrade, Da Silva, Ferreira, González, Carrera, 2004) con la correlación positiva o negativa entre los eventos.

Utilizando los métodos Rugosidad/longitud y Rango re-escalable en Benoit® para analizar la información obtenida en cada estado de los cuales se hizo este trabajo, los resultados se muestran en la Tabla 1, que muestra los valores obtenidos para el coeficiente de Husrt (H), dimensión fractal (D) y desviación estándar (DE) de cada estado:

Tabla 1.
Tabla de valores de coeficiente de Hurst, dimensión fractal y desviación estándar de los estados; Chiapas, Guerrero y Oaxaca

Estado / Método	Chiapas			Guerrero			Oaxaca		
	Coeficiente de Hurst	Dimensión fractal	Desviación estándar	Coeficiente de Hurst	Dimensión fractal	Desviación estándar	Coeficiente de Hurst	Dimensión fractal	Desviación estándar
Rango re-escalado	-0.001	2.001	0.1300285	-0.004	2.004	0.0467813	-0.007	2.007	0.0929217
Rugosidad longitud	0.050	1.950	0.0038225	0.033	1.967	0.0043215	0.041	1.959	0.0032110

Fuente: Elaboración propia de los autores 2016.

Evaluación e interpretación de los datos

De los resultados obtenidos en la tabla 1 con los métodos R/S y RL se obtuvo para los tres estados un valor de H entre 0 y 0,5 el cual indica la presencia de un proceso ergódico o anti persistente (*mean reverting* o ruido rosa). Esto indica que los movimientos hacia arriba serán sucedidos por movimientos hacia abajo (y vice-versa), los valores futuros tenderán a concentrarse en torno a una media de largo plazo (el proceso tiene una memoria larga). La fuerza de esta tendencia (de reversión a la

media) será mayor cuanto más cercano a 0 sea el exponente Hurst. Se considera que esta serie tiene ruido rosa, que es común en la naturaleza y está relacionado con procesos de relajación (equilibrio dinámico) y turbulencia.

En tanto que la Dimensión Fractal (D) es el número que refleja la medida topológica de un conjunto fractal a escalas distintas, esto es, definir como el número que sirve para cuantificar el grado de irregularidad y fragmentación de un conjunto geométrico o de un objeto natural (Strecker, 2004). Si D posee valores elevados (1.6 ≥ D ≥2), la serie confirma señales de patrones de tipo Bandas de Bollinger, RSI (Índice de Fuerza Relativa), *stochastics* y *reversals* (reversiones). Este tipo de activos poseen un comportamiento muy volátil (cómo lo indica su elevada dimensión fractal) y son ideales para *trading* activo de corto plazo. Y cómo es posible observar en los tres casos con los dos métodos el valor de D es cercano a 2, lo que indica que esta serie es muy volátil y anti persistente.

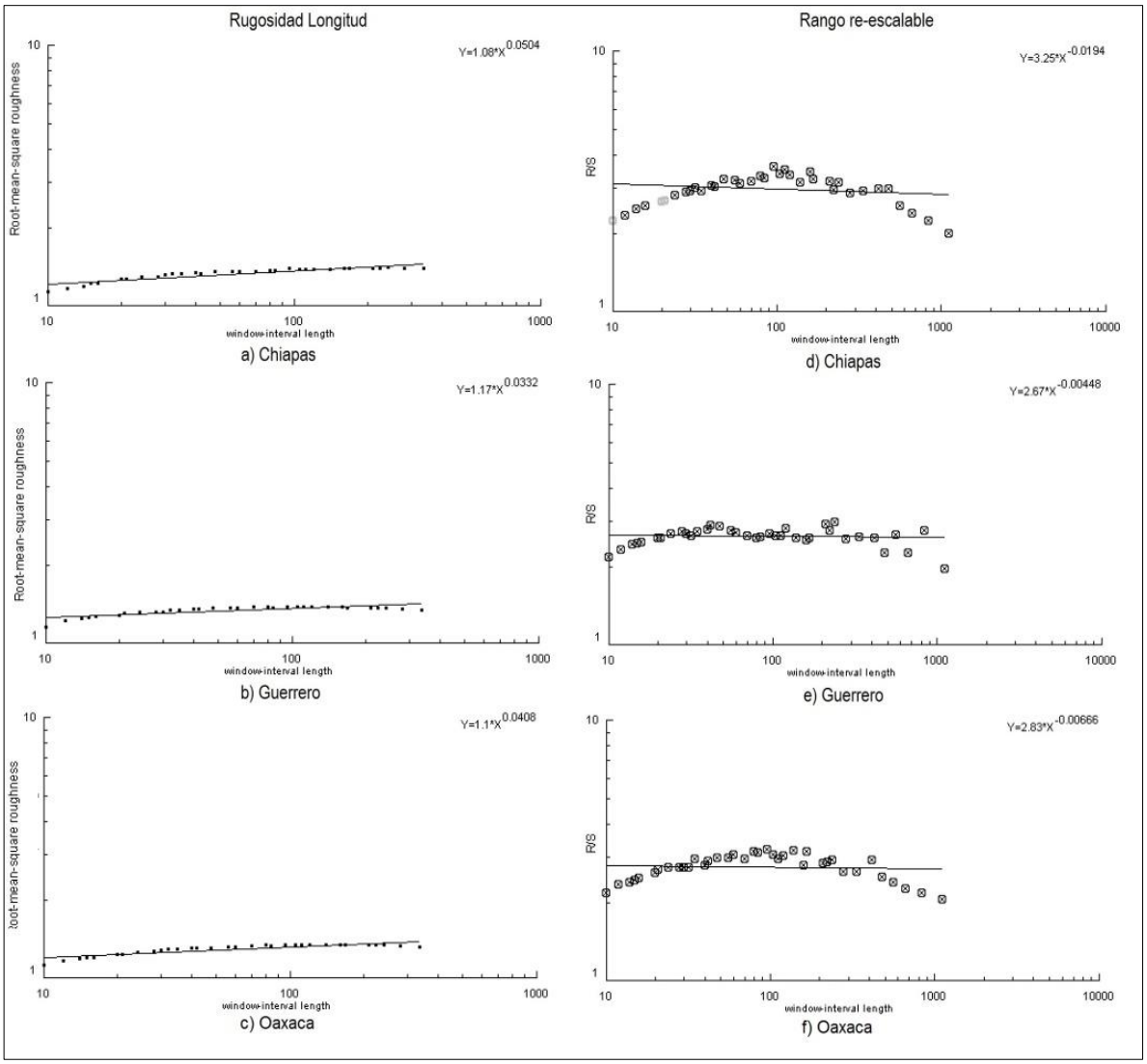
Los parámetros fractales, que estiman la tendencia o rugosidad de la distribución de los eventos sísmicos, mostraron que las series y los periodos de tiempo estudiados tienden a ser anti persistentes (0= H= 0.5) o de alta rugosidad, debido a la correlación negativa entre los eventos. La dimensión fractal mostro un resultado muy cercano a 2, lo cual confirma que las series de tiempo mensuales muestran tendencia a ser anti persistentes, que se relaciona con el tipo de eventos que se presentan en las regiones, las cuales se caracterizan por la alta variabilidad en magnitud, pero con mayor frecuencia; lo que se refleja en la rugosidad de la serie de tiempo. Tal como se menciona en Domínguez y Garzón (2011) la geometría fractal permite describir estructuras y procesos que ocurren en la naturaleza, ya que estos objetos tienen un alto grado de irregularidad, tal como es el caso de estudio de este trabajo en los sismos en México, también en Quezada (2006 y 2005), se habla de fractales como un intento de descubrir el comportamiento de los fenómenos naturales y sustenta que estas dimensiones fractales no son enteras y tienen muchas irregularidades.

Otro resultado arrojado por Benoit® son las gráficas de ajuste que se muestran en la figura 2.

RESULTADOS EXPERIMENTALES

Al analizar los resultados obtenidos en este trabajo, claramente se puede apreciar que el coeficiente de Hurst para los tres estados sobre los cuales se trabajó, es mayor que 0 y menor que 0.5, lo cual corroborándolo con lo que se sustenta en (Quintero y Ruiz, 2011; Rodríguez, 2014; Luengas, *et al.*, 2010) corresponde a un rudo rosa, es decir, que los datos pasados no tienen relación con los sucesos que pasaran, por lo cual el sistema es inestable y anti persistente, este tipo de valores para el coeficiente de Hurst suele presentarse en los fenómenos de la naturaleza, tal como es el caso de este trabajo.

Figura 2.
Gráficas de ajuste de los métodos rugosidad longitud y rango re-escalado en los tres estados



Fuente: Elaboración propia de los autores 2016.

CONCLUSIONES

La geometría fractal permite describir estructuras y procesos que ocurren en la naturaleza, ya que estos objetos tienen un alto grado de irregularidad, tal como es el caso de estudio de este trabajo acerca los sismos en México, también se habla de fractales como un intento de descubrir el comportamiento de los fenómenos naturales y sustenta que estas dimensiones fractales no son enteras y tienen muchas irregularidades. Al analizar los resultados obtenidos en este trabajo, claramente se puede apreciar que el coeficiente de Hurst para Chiapas, Guerrero y Oaxaca, es mayor que 0 y menor que 0.5, lo que corresponde a un rudo rosa, es decir, que los datos pasados no tienen relación con los

sucesos que pasaran, por lo cual, el sistema es inestable y anti persistente, este tipo de valores para el coeficiente de Hurst suele presentarse en los fenómenos de la naturaleza, aunado a esto, es posible estimar el valor de la dimensión fractal (D) mediante los valores ya mencionados del exponente de Hurst (H), que se calcula realizando la operación 2 – H, una vez teniendo el valor D para los tres estados de estudio, se puede apreciar que el valor es muy cercano a 2, esto nos indica que el sistema es muy irregular, ya que este valor D nos indica el grado de irregularidad de un sistema, y es más irregular cuando su valor D es más cercano a 2, como el caso de estudio de este trabajo.

BIBLIOGRAFÍA

Adeli, H., Panakkat, A. (2009*). A probabilistic neural network for earthquake magnitude prediction*. Neural Networks.

Aguirre, J. et al. (2015).”Análisis de deserción escolar con minería de datos”. *Revista Research in Computing Science* 93, pp. 1-2.

Balankin, A., Morales, O., Gálvez, M. (2004). “Crossover from antipersistent to persistent behaviour in time series possessing the generalized dynamic scaling law”. *Review Ser.* 69(3):45-54.

Berson, A., Smith, S. (1997). *Data Warehouse, Data Mining & OLAP.* Mc graw hill, USA.

Botero, S., Cano, J. (2008). “Análisis de series de tiempo para la predicción de los precios de la energía en la bolsa de Colombia”, *Revista Cuadernos de Economía,* volumen 27, número 48, pp. 18.

Burgos, T., Pérez, E. (1999). “Estimation of the fractal dimension of a rainfall time series over a zone relevant to the agriculture in Havana. SOMETCUBA”. *Bulletin.* Vol. 5. Num. 1. 35 p.

Cáceres, G., Rodríguez, J. (2011). “Agrupamiento de datos de series de tiempo. Estado del arte”, *Revista Vínculos*. vol. 8. No.1.

Dominguez, A., Garzon, D. (2011). “Comportamiento fractal espacial en la expansión de la distribución del flujo sanguíneo cerebral en Alzheimer”, *Revista cubana de investigaciones biomédicas*, Vol. 30, No. 3, pp. 2-4.

Espinoza, J., Jiménez, Z. (2016). *Terremotos y ondas sísmicas.* Marzo [En línea]. Disponible en: http://www2.ssn.unam.mx:8080/website/jsp/Cuaderno1/ondas-index.html

Fayyad, U., Piatetsky-Shapiro, et al. (1996). “The KDD process for extracting useful knowledge from volumes of data.” *Comunications of the ACM* Vol 39, No 11, New York (USA): ACM Digital Library p 27-34 ISSN: 0001-0782.

Frawley, W., Piatetsky-Shapiro, Matheus, C. (1992). “Knowledge discovery in databases: An Overview” *AI magazine*, Vol 13, No 3, pp 57.

Gerardo, R., Zenón, J. (2016). *Sismos en la ciudad de México y el terremoto del 19 de septiembre de 1985,* Marzo [En línea]. Disponible en: http://www2.ssn.unam.mx:8080/website/jsp/SISMO85-4.HTM

Hand, D., Mannila, H., Smyth, P. (2001). *Principles of Data Mining* Cambridge, MA: The MIT Press.

Hernández, J., Ramírez, M., Ferri, C. (2004), *Introducción a la Minería de Datos*. Pearson Educación, S.A.

Kagan, Y., Jackson, D. y Rong, Y. (2007). *A testable five-year forecast of moderate and large earthquakes in southern California based on smoothed seismicity.* Seismological Research Letters.

Luengas, D., et al. (2010). “Metodología e interpretación del coeficiente de Hurst”, *Revista Odeon,* número 5, pp. 10.

Mandelbrot, B. (1988). *Los objetos fractales: forma, azar y dimensión*. España: TusQuets Editores.

Mandelbrot, B. (2002). *Gaussian self-affnity and fractals. Globality. The Earth, 1/f Noise, and R/S. Springe*r 654 p.

Martínez, Á, Troncoso, A., Morales, E., Riquelme, J. (2011). *Computational intelligence techniques for predicting earthquakes. Lecture Notes in Artificial Intelligence*.

Miranda, J., Andrade, R., Da Silva, A., Ferreira, C., González, A., Carrera, J. (2004). “Temporal and spatial persistence in rainfall records from Northeast Brazil and Galicia”. Theor. Appl. Climatol. 77:113-121.

Morales, E., Martínez, Á., Troncoso, A., De Justo, J., Rubio, C. (2010). *Pattern recognition to forecast seismic time series. Expert Systems with Applications.*

Palomas, M. (2002). “Evidencia e implicaciones del fenómeno Hurst en el mercado de capitales”. *Gaceta de Economía.* Año 8. 15:117-153.

Quezada, A. (2005). “Fractales más allá de 1D, 2D O 3D,” *Revista digital universitaria*, volumen 6, número 12, pp. 4-6.

Quezada, A. (2006). “Fractales en el estudio de la psicología”, *Revista digital universitaria,* volumen 7, número 10, pp. 4.

Quintero, O., Ruiz, J. (2011). “Estimación del exponente de Hurst y la dimensión fractal de una superficie topográfica a través de la extracción de perfiles”. *Revista geomática* ud. Geo, pp. 2-4

Riquelme, J., et al. (2006). “Minería de Datos: Conceptos y Tendencias*”. Revista Iberoamericana de Inteligencia Artificial*, volumen 10, número 29, pp. 1-3.

Rodríguez, J. (2006). *Clasificación de series de tiempo por minería de datos.* Tesis M. SC., Instituto Politécnico Nacional. México D.F.

Rodríguez, J. (2008). “Minería de datos para la determinación del grado de exclusión social”, *Revista Vínculos*, volumen 5, número 1, pp. 2.

Rodriguez, R. (2014).”El coeficiente de Hurst y el parámetro α-estable para el análisis de series financieras. Aplicación al mercado cambiario mexicano”, *Revista Contaduría y administración*, volumen 59, número 1, pp. 6-7.

Salas, J., Delleur, J., Yevjevich, V., Lane, W. (1985). *Applied modeling of hydrologic time series. Water Resources Publications*. Littleton, CO. USA.

Schroeder, M. (1991). *Fractals, chaos, power laws: Minutes from an infinite paradise*. Freeman, W. H. & Co. New York. New York. USA.

Simon, A. (1997). *Data Warehouse, data Mining and OLAP.* John Wiley & Sons USA.

Strecker, J. (2004). *Fractional Brownian Motion Simulation: Observing Fractal Statistics in the Wild and Raising Them in Captivity*. Wooster: The College of Wooster, Department of Mathematics and Computer Science.

Timarán, R., Calderón, A., Jiménez, J. (2013). “Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos”, *Revista Vínculos*, volumen 10, número 1, pp. 5.

Velarde, A. (2003). “Minería de datos: una introducción”. *Revista Ciencia tecnológica*, número 23, pp. 1-2.

White, C. (2001). *IBM Enterprise Analytics for the Intelligent e-business.* IBM Press USA.

Zúñiga, F. (2016). *Predicción sísmica*, Marzo [En línea]. Disponible en: http://www2.ssn.unam.mx:8080/website/jsp/Prediccion/ramón.jsp.