Ciencias de la Tierra y el Espacio, julio-diciembre, 2022 Vol.22, No. 2, pp. 50-62 ISSN 1729-3790

Predicción de valores faltantes del índice geomagnético K medido en Cuba mediante técnicas estadísticas y minería de datos

Ernesto Alejandro Piloto-Zaldívar, Omar González-Trujillo, Oscar Alejandro Sánchez-Sánchez, Zulia Bárbara Bárcena-Fonseca

Instituto de Geofísica y Astronomía, Departamento de Geofísica Espacial, Cuba, piloto@iga.cu
ernestoalejandropiloto@gmail.com, ssaoscar13@gmail.com, omarito@iga.cu, barcena@iga.cu.
ORCID: 0000-0003-0162-7529.

Recibido: Junio 2022	Aceptado: noviembre 2022

Resumen

Se definió el modelo a emplear para obtener un registro más completo de los valores del índice geomagnético K, que no pudieron ser medidos por el observatorio geomagnético del Instituto de Geofísica y Astronomía tomando como muestra los datos recogidos durante los años 2006, 2007 y 2008. Se confeccionó una base de datos para determinar la relación existente entre el índice K medido en Cuba y el índice planetario Kp. Se usó el software Matlab para conocer la linealidad existente entre estos índices, dando como resultado un modelo de regresión lineal ajustado a los datos, que permitirá predecir los valores faltantes del índice K. Se entrenó una red neuronal, a partir del algoritmo de perceptrón multicapa con retropropagación para obtener un modelo de predicción, utilizando el software Weka. Se compararon los resultados obtenidos al utilizar las técnicas estadísticas y la minería de datos.

Palabras clave: índice geomagnético, minería de datos, predicción, regresión, técnicas estadísticas.

Prediction of missing values of the geomagnetic index K measured in Cuba using statistical techniques and data mining

Abstract

The objective of this work is to define the model to be used to obtain a more complete record of the values of the geomagnetic index K, which could not be measured by the geomagnetic observatory of the Institute of Geophysics and Astronomy, taking as a sample the data collected during the years 2006, 2007 and 2008. A database was created where it was carried out to determine the relationship between the K index measured in Cuba and the planetary Kp index. Matlab software was used to determine the linearity between these indices, resulting in a linear regression model adjusted to the data, which will allow predicting the missing values of the K index. A neural network was trained, based on the multilayer perceptron algorithm with backpropagation to obtain a prediction model, using the Weka software. The results obtained using statistical techniques and data mining were compared.

Keywords: data mining, geomagnetic index, magnetic storms, neural networks, prediction, regression, statistical techniques.

1. Introducción

El IGA es la institución líder en las investigaciones referentes al geomagnetismo y en especial, los aportes de estudios sobre el comportamiento de la ionosfera en Cuba. Desde el año 2006 no se realizan correctamente tales contribuciones, ya que desde ese momento el equipo con que se trabaja en el observatorio empezaba a presentar dificultades para realizar las mediciones, debido a la obsolescencia tecnológica existente en el centro, además de no contar con suficiente personal con la experiencia para desempeñar estas actividades.

A esta problemática se le suma que, en el año 2009, en las cercanías del IGA, ubicado en el municipio La Lisa, La Habana, se construyeron edificios de viviendas de lo que hoy se conoce como comunidad Granma. Estas construcciones afectaron al observatorio geomagnético que se encontraba en el instituto, debido a que aumentaron la señal de ruido ambiental y hacía imposible realizar su trabajo, a pesar de las correcciones de datos que se hacían para evitar que esto sucediera.

Esto implicó que en años posteriores fuera desmontado el observatorio y se dejaran de obtener valores del índice K, y que el índice Kp, medido a partir de una red global de observatorios, no se obtuviera con exactitud. Sin embargo, para la obtención de este último se usaron los datos disponibles medidos por la red de observatorios geomagnéticos cercana a la región de Cuba, los cuales poseían un margen de error, al no contar con los datos cubanos.

La predicción del índice K es importante para aplicaciones como la aeronáutica, topografía, orientación de los campos de antena, ya que permite pronosticar las tormentas magnéticas y el estado de la ionósfera para las comunicaciones por circuitos de onda corta. También es importante tener un registro del comportamiento de este índice que sea fiable, en caso de que se requiera hacer un estudio histórico de las afectaciones vinculadas a tormentas magnéticas en el país.

El objetivo de este trabajo es definir el modelo a emplear para obtener un registro más completo de los valores del índice geomagnético K que no pudieron ser medidos por el observatorio geomagnético del IGA durante los años 2006, 2007 y 2008, y a partir de esto obtener un modelo de regresión utilizando técnicas estadísticas que permita predecir los valores faltantes del índice K medido en Cuba durante los años 2006, 2007 y 2008, así como obtener un modelo de predicción utilizando minería de datos para hallar los valores faltantes del índice K medido en Cuba durante los años 2006, 2007 y 2008 y comparar el resultado de ambas técnicas empleadas.

Los resultados alcanzados en la investigación fueron:

- Índices de correlación de Pearson, Spearman y Kendall para los valores de los índices geomagnéticos K y Kp correspondientes a los años 2006, 2007 y 2008.
- Espectros de Welch y Lomb-Scargle para los valores de los índices geomagnéticos K y Kp correspondientes a los años 2006, 2007 y 2008, que permitieron conocer fenómenos que se manifiestan en el dominio de la frecuencia, no siendo así en el dominio del tiempo
- Modelo de regresión polinómica lineal para los valores totales del índice K a partir de los valores totales del Kp para los años 2006, 2007 y 2008.
- Modelo de red neuronal a partir del algoritmo de perceptrón multicapa con propagación hacia atrás para los valores totales del índice K a partir de los valores totales del Kp para los años 2006, 2007 y 2008.
- Comparación de la efectividad de las técnicas estadísticas y minería de datos para la predicción de valores faltantes del índice K durante los años 2006, 2007 y 2008.

-

2. Materiales y métodos

2.1 Análisis de secuencia de datos

Las funciones de autocovarianza, covarianza cruzada, autocorrelación y correlación cruzada son empleadas con el propósito de comparar series temporales y revelar sus estructuras internas. Todas estas funciones se emplean mucho en el procesamiento de datos geofísicos, tales como los sísmicos, geoeléctricos, electromagnéticos (especialmente los de georradar), magnéticos y gravimétricos.

El coeficiente de correlación de Pearson (r) es una medida de dependencia lineal entre dos variables aleatorias cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.

El coeficiente de correlación de Spearman, ρ (ro) es una medida de la correlación (la asociación o interdependencia) entre dos variables aleatorias (tanto continuas como discretas). Para calcular ρ, los datos son ordenados y reemplazados por su respectivo orden.

El coeficiente de correlación de rango de Kendall, comúnmente conocido como coeficiente τ de Kendall (con la letra griega τ, tau), es una estadística utilizada para medir la asociación ordinal entre dos cantidades medidas. Una prueba τ es una prueba de hipótesis no paramétrica para la dependencia estadística basada en el coeficiente τ.

2.2 Análisis espectral de los datos

El estudio de los espectros constituye una herramienta fundamental para la comprensión del comportamiento de las funciones en el dominio de la frecuencia, lo que propicia el conocimiento de ciertos fenómenos que no se manifiestan claramente en el dominio del tiempo, y está enfocado al estudio de los pasos para la estimación de espectros de señales observadas o medidas empíricamente, y de las que en general se desconoce su expresión analítica.

- Estimado de la densidad espectral de potencia del periodograma. Pxy = periodogram(x). Estima la densidad espectral de potencia (DEP, PSD) de una señal discreta. La señal discreta x puede ser un vector o una matriz.
- Estimado de la densidad espectral de potencia del periodograma por el método de Lomb–Scargle. [pxx, f] = plomb (x, t). Devuelve la densidad espectral de potencia estimada por el método de Lomb – Scargle de una señal x que está muestreada no uniformemente en los puntos especificados por t, que tiene que crecer monótonamente, pero no uniformemente espaciados. Todos los elementos de t tienen que ser no negativos. La señal x puede ser un vector o una matriz. Tanto x como t pueden contener valores NaN (not a number), que son tratados por el algoritmo como ausencia de datos, y excluidos de los cálculos (en este trabajo los valores NaN corresponden a los datos faltantes del índice K que no pudieron ser medidos por distintas razones en los años 2006, 2007 y 2008). Si no se tiene el vector t, se pueden incluir valores NaN directamente en la señal x. Esto evita que cambien las frecuencias normalizadas a la salida. La variable de salida pxx es evaluada para las frecuencias que se devuelven en la variable de salida f.
- Estimado de la densidad espectral de potencia por el método de Welch. pxx=pwelch(x). Devuelve la densidad espectral de potencia (PSD) de la señal de entrada x usando el estimador de promediación de segmentos superpuestos de Welch. La señal x puede ser un vector o una matriz. Por defecto, x es dividida en las secciones más largas posibles, para obtener lo más cercano posible, pero no exceder a 8 segmentos con 50 % de superposición. Cada sección se somete a la acción de una ventana de Hamming.
- Espectrograma usando la transformada de Fourier de corto tiempo. spectrogram(x). El espectrograma es un gráfico que relaciona las frecuencias con los tiempos, por lo que permite evaluar el comportamiento de las frecuencias instantáneas y las amplitudes de sus componentes en el tiempo. La señal de entrada x es un vector.
-

2.3 Detección de transiciones abruptas en series de tiempo

En la mayoría de los casos, las tendencias y eventos en los dominios de tiempo y frecuencia se detectan calculando los parámetros estadísticos de los datos (por ejemplo, medidas de tendencia central y dispersión) contenidos en una ventana deslizante de longitud L. La precisión de estos parámetros depende de la longitud de la ventana, es decir, un valor exacto para la media y la varianza se obtiene si L es grande.

Las estadísticas clásicas de la prueba t y la prueba F se utilizan a menudo para comparar las medias y las varianzas de dos conjuntos de medidas y, por lo tanto, podrían utilizarse para detectar cambios en la ubicación y la dispersión entre dos ventanas deslizantes. Estas dos pruebas, sin embargo, hacen el supuesto básico de que las muestras provienen de una población con una distribución gaussiana.

Las pruebas no paramétricas de Mann-Whitney y Ansari-Bradley proporcionan una solución a este problema que es independiente de la distribución. La prueba de Mann-Whitney realiza una prueba de suma de rangos bilateral de la hipótesis nula de que dos muestras provienen de distribuciones continuas idénticas con medianas idénticas, contra la alternativa de que no tienen medianas idénticas. La prueba de Ansari-Bradley realiza una prueba bilateral de que dos muestras independientes provienen de la misma distribución, contra la alternativa de que provienen de distribuciones que tienen la misma mediana y formas, pero diferentes dispersiones (Trauth, 2015).

Análisis estadístico de los datos

El análisis de los datos se realiza después del pre-procesamiento y antes de aplicar la minería de datos, con el fin de conocer la calidad de estos datos y saber si el resultado de aplicar la minería de datos es un resultado confiable.

Para esto se aplican varias pruebas de hipótesis entre las cuales están:

- Prueba de hipótesis t de Student para la media de una sola muestra: Esta prueba se usa para probar si la media de una muestra con distribución normal o gaussiana, posee un valor igual al valor mu especificado (hipótesis nula) o si los valores son diferentes (hipótesis alternativa). Esta es una prueba paramétrica, que se aplica solo para muestras con distribución normal, cuando se desconoce la desviación típica y se estima a partir de la muestra.
- Prueba de hipótesis t de Student para dos muestras: Esta prueba se usa para probar la hipótesis de que dos muestras extraídas de una misma población con distribución normal o gaussiana, tienen idénticas medias. Por ejemplo: se han obtenido dos muestras de n1 y n2 mediciones independientes presuntamente a partir de una misma población; por ejemplo, mediciones de dos conjuntos de muestras de rocas, tomadas de dos afloramientos diferentes, pero de un mismo tipo de roca. La prueba t de Student para dos muestras es una prueba paramétrica. La prueba t puede ser usada para determinar si ambas muestras proceden de una misma población; es decir, de la misma unidad litológica (hipótesis nula) o de dos poblaciones diferentes (hipótesis alternativa). Ambas distribuciones muestrales tienen que ser normales o gaussianas, y las varianzas para los dos conjuntos de mediciones deben ser similares. La prueba t permite establecer inferencias acerca de las diferencias entre las medias de dos muestras.
- Prueba de hipótesis de Mann - Whitney para dos muestras o Prueba de suma de rangos de Wilcoxon: Esta prueba se usa para determinar si dos muestras provienen de la misma población, por ejemplo, la misma unidad litológica (hipótesis nula), o de dos poblaciones diferentes (hipótesis alternativa). En contraste con la prueba t de Student, que compara las medias de datos con distribución normal o gaussiana, esta prueba de Mann - Whitney compara las medianas sin exigir una suposición de normalidad para la población subyacente; es decir, esta es una prueba de hipótesis no paramétrica. Esta prueba requiere que ambas muestras tengan dispersiones o desviaciones típicas similares (pero no necesariamente iguales).
- Prueba de hipótesis de Kolmogorov - Smirnov para una muestra y una distribución teórica de probabilidades: Esta prueba se usa para determinar si la muestra proviene de una población con determinada distribución teórica de probabilidades. Por ejemplo, una unidad litológica con determinada distribución teórica de probabilidad conocida (hipótesis nula). Esta prueba no exige que la muestra o la población tengan una distribución normal o gaussiana de probabilidades. Es una prueba no paramétrica. Puede ser empleada con magnitudes medidas solamente en escalas de intervalos o de proporciones.
- Prueba de hipótesis de Kolmogorov - Smirnov para dos muestras: Esta prueba verifica la igualdad o no de las distribuciones de probabilidad correspondientes a dos muestras. Es una prueba no paramétrica porque no exige que las muestras o la población tengan una distribución normal o gaussiana de probabilidades. Sirve para determinar si las dos muestras provienen de una misma población (o al menos de dos poblaciones con iguales distribuciones de probabilidades arbitrarias). Por ejemplo, si las muestras se tomaron de una misma unidad litológica (hipótesis nula) o si las dos muestras proceden de dos poblaciones con diferentes distribuciones de probabilidad (hipótesis alternativa).
- Prueba de hipótesis F de Snedecor y Cochran para dos muestras: La prueba de hipótesis F de Snedecor y Cochran para dos muestras compara las varianzas Sa2 y Sb2 de dos distribuciones, donde Sa2 > Sb2. La prueba presupone que la varianza de la población de donde se extrajo la muestra 1 es igual a la varianza de la población de donde se extrajo la muestra 2 (hipótesis nula). La hipótesis alternativa es que ambas varianzas son desiguales. Un ejemplo de su uso es la comparación de la heterogeneidad natural de dos muestras sobre la base de mediciones replicadas. Los tamaños de las muestras na y nb deben ser ambos mayores que 30. Tanto las distribuciones de las muestras como de las poblaciones tienen que ser normales o gaussianas; es decir, la prueba F de Snedecor y Cochran es una prueba paramétrica.
-

2.5 Regresión lineal y análisis de residuos

La regresión lineal ofrece otra forma de describir la relación entre las dos variables x e y. Mientras que el coeficiente de correlación de Pearson proporciona solo una medida aproximada de una tendencia lineal, los modelos lineales obtenidos por análisis de regresión permiten la predicción de valores y arbitrarios para cualquier valor dado de x dentro del rango de datos. Las pruebas estadísticas de significancia del modelo

lineal proporcionan algunas ideas sobre la precisión de estas predicciones. La regresión clásica asume que, y responde a x, y que toda la dispersión en el conjunto de datos está contenida dentro del valor de y (figura 1).

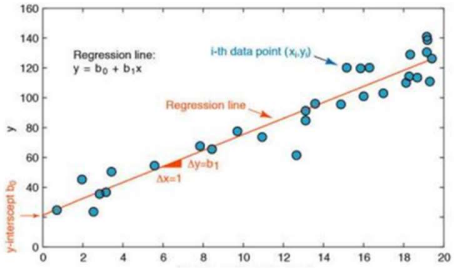


***Fig. 1**: Regresión lineal. Mientras que la regresión clásica minimiza las desviaciones Δy, la regresión reducida del eje principal minimiza el área triangular 0,5*(ΔxΔy) entre los puntos de datos y la línea de regresión, donde Δx y Δy son las distancias entre los valores x e y predichos y verdaderos. La intersección de la línea con el eje y es b0 y la pendiente es b1. Estos dos parámetros definen la ecuación de la recta de regresión (Trauth, 2015).*

Método de minería de datos

La sinapsis tiene un potencial o intensidad ajustable, y se cree que esta cualidad o adaptabilidad, entre otras causas, esté en la base de lo que se conoce como aprendizaje, siendo por ende esta propiedad la que fundamenta una red neuronal artificial o sistema de ellas, por cuanto la base matemática primordial la constituye el ajuste de las sinapsis entre las neuronas (García, 2005).

Desde el punto de vista de su estructura la red neuronal (figura 2) consta de:

- Un grupo de conectores caracterizados por un peso o fuerza propia. Específicamente una señal xj como una entrada de sinapsis j, conectada a una neurona, es multiplicada por un peso sináptico Wkj. El primer subíndice 12 de los pesos sinápticos hace referencia al número de la neurona y el segundo a la entrada de la sinapsis a la cual se refiere el peso.
- Un sumador de señales de entrada, tomado en cuenta los pesos por la respectiva sinapsis de la neurona.
- Una función de activación para limitar la amplitud de la salida de una neurona. La función de activación limita el rango de valores permisibles de amplitud de la señal de salida a algún valor finito.
-

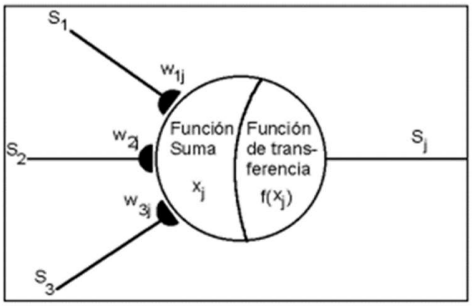


***Fig. 2**: Modelo de una red neuronal (Catalina, 1994)*

En el presente trabajo se utilizó el software Weka, el cual tiene varias herramientas para la minería de datos, dentro de las cuales está la red de tipo perceptrón multicapa, que aprovecha la naturaleza paralela de las redes neuronales para reducir el tiempo requerido por un procesador secuencial para determinar la salida adecuada a partir de una entrada. Además, el tiempo de desarrollo se reduce como consecuencia de que la red puede aprender el algoritmo correcto sin que alguien tenga que deducir por anticipado el algoritmo en cuestión.

La red perceptrón multicapa es entrenada de forma supervisada mediante el algoritmo de retropropagación, esta es la razón por la que también se conoce como red de retropropagación o red de propagación hacia atrás. Al hablar de red de retropropagación o red de propagación hacia atrás se hace referencia al algoritmo de aprendizaje y no a la propagación de las señales durante el funcionamiento de la red. La retropropagación consiste en propagar el error hacia atrás durante el entrenamiento, es decir, de la capa de salida hacia la capa de entrada, pasando por las capas ocultas (Castro, 2006).

3. Resultados y discusión

Para el procesamiento de datos fue necesario agrupar los valores de los índices K y Kp separados por año en una base de datos en Excel, para así poder identificar cuáles son los años en que se tienen los valores de los dos índices y cuáles no.

Al analizar la base de datos ya creada se observó que en los años en que se tenían mediciones de los dos índices (años 2006, 2007 y 2008) faltaban valores del índice K. Esto se debe a varios factores que propiciaron que la estación no estuviera realizando mediciones en ese momento como son, caída momentánea de la red eléctrica, tormentas tropicales o incluso mantenimiento a la misma estación.

Estos valores faltantes fueron tomados en la base de datos como NaN (Not a Number). Para aplicar las técnicas de estadísticas y minería de datos como las redes neuronales descritas en el apartado anterior se hizo una normalización de los datos mediante el software MATLAB, en el cual se restó la media de cada dato y se dividió entre su desviación estándar. De esta forma los valores NaN tomaron valor 0.

Una vez normalizados los datos se aplicaron técnicas de autocorrelación y correlación cruzada a los valores de K y Kp correspondientes a cada época, es decir, cada K con cada Kp del mismo año.

Como resultado se obtuvo que la mejor correlación fue entre los valores del índice K y Kp del 2007 (figura 3), ya que este es el año que menos valores NaN tiene.

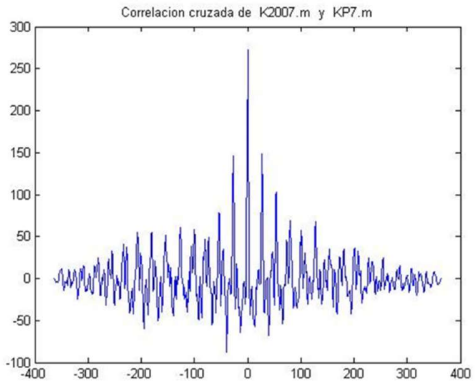


***Fig. 3**: Correlación cruzada entre los valores de índices K y Kp del año 2007*

Se realizó el cálculo de los coeficientes de Pearson, Spearman y Kendall de los valores de los índices K y Kp para cada año, cuyo resultado se muestra en la siguiente tabla:

Tabla I. Resultado de los coeficientes de Pearson, Spearman y Kendall.

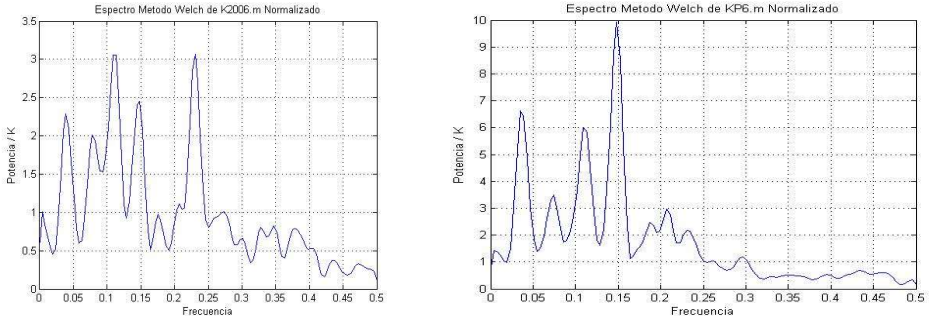
Coeficientes	Año 2006	Año 2007	Año 2008
R de Pearson	0.5897	0.7563	0.6604
Tau de Spearman	0.5103	0.7131	0.5686
Rho de Kendall	0.3872	0.5409	0.4205

Como se puede observar en la tabla 1 existe una correlación positiva entre los índices K y Kp. El año 2007 muestra mayores valores de correlación indicando una mejor asociación entre los datos. Esto puede ser debido a la escasa cantidad de valores NaN, que, aunque las tres series (2006, 2007 y 2008) fueron normalizadas, por ser tomados como 0, siguen afectando el comportamiento de estas series.

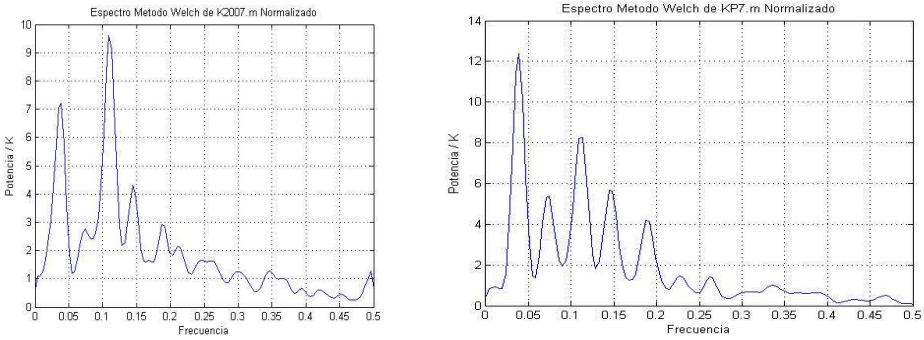
Para la comprensión del comportamiento de los índices de K y Kp en el dominio de la frecuencia, se realizó el estudio de los métodos de espectro de Welch, Lomb-Scargle, periodograma y espectrograma.

A partir del análisis de estos espectros se observó que el método de espectrograma no aportó mucha información en cuanto al comportamiento en el dominio de la frecuencia ya que dio como resultado un gráfico cuya interpretación resultó ser demasiado compleja. Sin embargo, los métodos de Welch y Lomb-Scargle fueron los que arrojaron mejores resultados ya que su lectura resulto ser mucho más asequible.

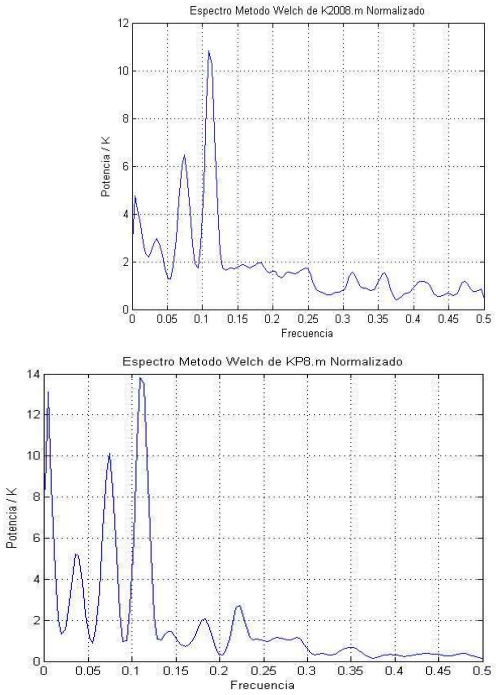
A partir del método Welch se observó una cierta correspondencia del índice K y Kp en los años 2007 (figura 4b) y 2008 (figura 4c), siendo menor en el 2006 (figura 4a).



(a)



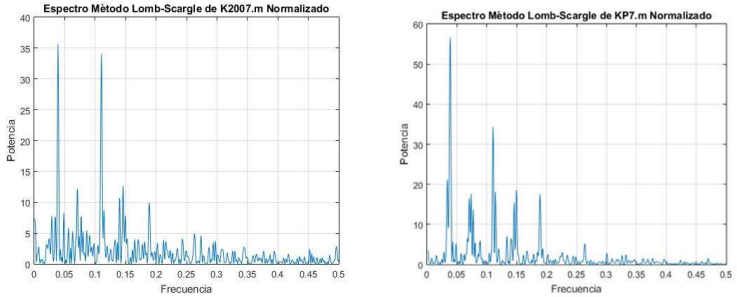
(b)



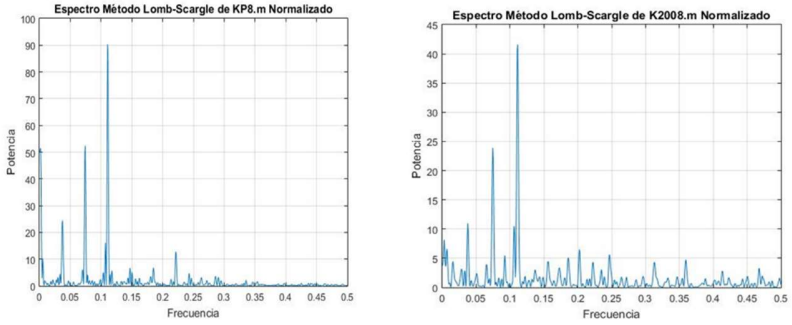
(c)

Fig. 4*: Método de espectro de Welch de los índices K y Kp para cada año: (a) 2006; (b) 2007; (c) 2008.*

El método de Lomb-Scargle permitió observar la ocurrencia de un fenómeno en el año 2007 (figura 5a) que se vio reflejado en el índice K, no siendo así en el índice Kp donde se refleja con menor potencia. Donde existe mayor correspondencia es en el año 2008 (figura5b) donde los dos gráficos tienen prácticamente el mismo comportamiento.



(a)



(b)

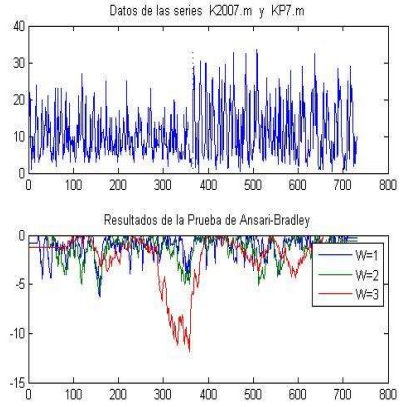
Fig. 5*: Método de espectro de Lomb-Scargle de los índices K y Kp para: (a) 2007; (b) 2008.*

Para detectar cambios abruptos en los valores de los índices K y Kp se tomaron estos datos como series en el tiempo, y se realizaron las pruebas de hipótesis de Mann-Whitney y Ansari-Bradley para conocer el momento exacto en que varía la mediana, forma y dispersión de las series, respectivamente.

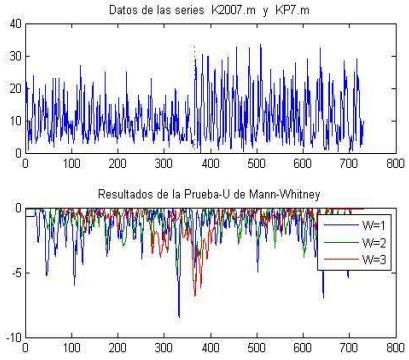
El mejor resultado se dio al analizar el año 2007 (figura 6), debido a que en los otros años se observaron varias transiciones, consecuencia de la calidad de los datos con los que se cuenta (existencia de valores NaN),

La prueba de hipótesis de Ansari-Bradley (figura 6a) resultó ser inefectiva para una ventana de mayor longitud, ya que mostraba una transición del valor de la dispersión para un tiempo de duración mayor que para las otras dos ventanas menores; al contrario de la prueba de Mann-Whitney (figura 6b) que solo lo hacía para un período de tiempo menor, pero que correspondía a la misma ubicación en el tiempo (entre los meses de abril y julio, y alrededor del mes de octubre).

La otra transición abrupta que ocurre es después del día 365 o alrededor de este y no se tuvo en cuenta a la hora de su análisis, ya que pertenece a la transición de la serie temporal de K y la del Kp, puesto que una está a continuación de la otra, pero pertenecen al mismo período de tiempo.



(a)



(b)

***Fig. 6**: Análisis de las transiciones abruptas de los índices K y Kp para el año 2007: (a) Utilizando la prueba de Ansari-Bradley; (b) Utilizando la prueba de Mann-Whitney.*

Al realizarse el estudio de las pruebas de hipótesis para ambos índices, estos no fueron tratados como series temporales, sino más bien, a partir de este momento en el procesamiento, como simples conjuntos de datos. Al no conocerse la distribución de probabilidad de los datos, se tomaron como no paramétricos, por lo que las pruebas que mejor reflejan la comparación entre los dos conjuntos de datos son las pruebas de Mann-Whitney y Kolmogorov-Smirnov, ya que son pruebas no paramétricas, es decir, no exige una suposición de que las muestras o poblaciones posean una distribución normal o gaussiana de probabilidades.

Las distintas pruebas de hipótesis permitieron llegar a las siguientes conclusiones de que las pruebas de Kolmogorov-Smirnov y Mann-Whitney coinciden en sus resultados en que los datos del 2006 no proceden de una o dos poblaciones con igual distribución de probabilidades, o que si proceden de dos poblaciones con la misma distribución pero que sus parámetros (dígase media, dispersión, mediana) no son iguales. Mientras que para el año 2007, ambas pruebas coinciden en que si proceden de la misma población o de dos con iguales distribuciones de probabilidades.

A partir de la regresión de los datos se obtuvieron dos ecuaciones de regresión de grado 1 para los tres años juntos (K_{tot} y Kp_{tot}). La ecuación (1) tiene como variable independiente a K_{tot} y la ecuación (2) a Kp_{tot}.

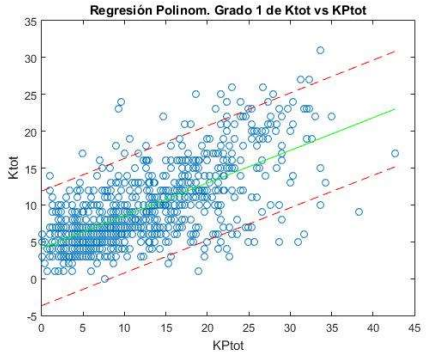
$$Kp_{tot} = K_{tot} + 2.77 \quad (1)$$

$$K_{tot} = 0.4\, Kp_{tot} + 4.1 \quad (2)$$

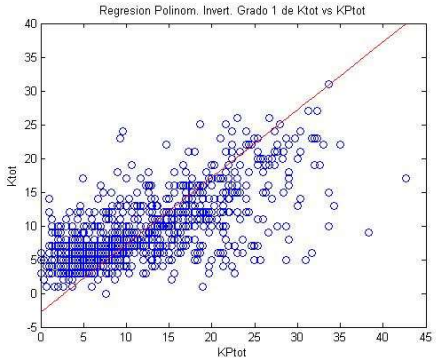
Puesto que lo que se necesita hallar es el K_{tot} a partir del Kp_{tot}, se hizo una inversión de variables a la ecuación (1) obteniendo así una nueva ecuación de regresión (3).

$$K_{tot} = Kp_{tot} - 2.77 \quad (3)$$

Para conocer la calidad de estos dos modelos de regresión se hizo un análisis de residuos a los modelos descritos por las ecuaciones (2) (figura 7a) y (3) (figura 7b).



(a)



(b)

***Fig.** 7: Modelos de regresión para Ktot vs Kptot: (a) modelo 2; (b) modelo 3.*

Al analizar los residuos se pudo observar una tendencia significativa en los correspondientes al modelo de regresión (3) (figura 8a), por lo cual se descartó este modelo, ya que esto indica que el modelo no describe completamente los datos; mientras que el modelo (2) (figura 8b) no muestra esta tendencia en sus residuos, por lo que se tomó como el modelo de regresión que describe los datos.

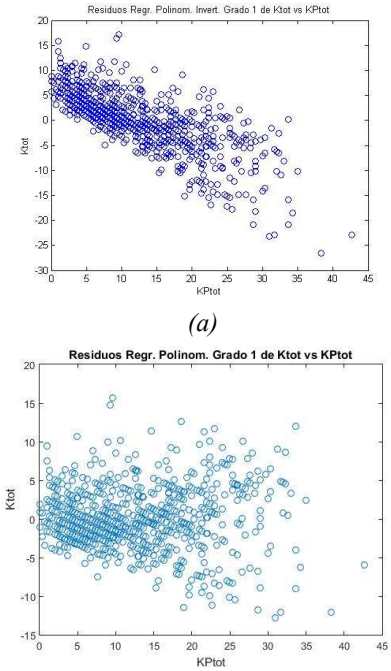


***Fig. 8**: Residuos del modelo de regresión 3*

Al aplicar la minería de datos se usó el software Weka, que dentro de sus herramientas tiene el modelo de perceptrón multicapa para las redes neuronales.

Se usó como red de entrenamiento la totalidad de los datos durante los años 2006, 2007 y 2008, eliminando los valores NaN del índice K y su respectivo valor Kp.

El procesamiento a partir de la red neuronal dio como resultado una señal de entrada y una señal de salida con su respectivo peso y una función de activación (figura 9).

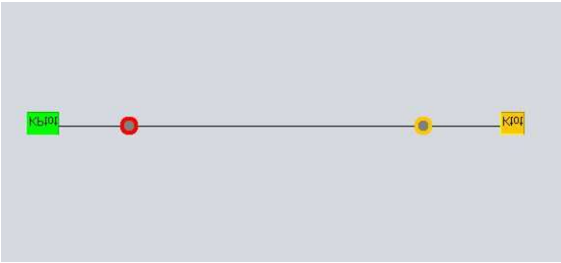


***Fig. 9**: Red Neuronal obtenida a partir de los datos de los años 2006, 2007 y 2008*

Como se puede observar, el modelo es muy sencillo y muy similar a lo que sería un modelo de regresión. Esto se debe a la cantidad de variables con las que se cuenta, ya que las redes neuronales se usan para modelos con grandes cantidades de señales de entrada.

Conclusiones

A partir del empleo de técnicas estadísticas se obtuvo un modelo de regresión lineal de una ecuación polinómica de grado 1, usando la totalidad de los datos útiles de la base de datos, exceptuando los valores NaN y sus respectivos valores del índice Kp, con este índice como variable independiente mediante el cual se puede obtener valores del índice K.

A partir del empleo del uso de la minería de datos, específicamente redes neuronales artificiales, se obtuvo un modelo de predicción mediante el algoritmo perceptrón multicapa con propagación hacia atrás. Se utilizó la totalidad de la base de datos, exceptuando los valores NaN y sus respectivos valores del índice Kp, para crear la red de entrenamiento que dio como resultado una nueva red con una señal de entrada, su peso, un sumador de señales de entrada y una señal de salida.

Al comparar el uso de ambas técnicas se obtuvo que la regresión estadística resultó ser más eficaz, ya que dio un resultado más amplio. Durante el procesamiento se pudo comprender mejor la relación entre los datos que componen la base de datos, y se observó la existencia de dos modelos de regresión con el índice Kp como variable independiente, de los cuales uno de ellos fue descartado al estudiar su calidad mediante el análisis de residuos. Mientras que la red neuronal tuvo mejor eficiencia, debido a que el resultado se obtuvo en menor tiempo, pero solo se obtuvo un modelo de predicción, sin dar características de este y sin saber la naturaleza de los datos, lo cual resultaba necesario para conocer la calidad del resultado. Aunque la estructura del modelo de regresión y de predicción es muy similar, la calidad y cantidad de los datos influyó mucho en los resultados, puesto que las redes neuronales normalmente trabajan con un gran volumen de variables independientes.

Referencias bibliográficas

Castro, J.F. 2006. *Fundamentos para la implementación de la red perceptrón multicapa mediante software*. Tesis en opción al grado de Ingeniero Informático, Universidad de San Carlos de Guatemala, Guatemala.

Catalina, G.A. 1994. *Introducción a las redes neuronales artificiales.*

García, M.G. 2005. *Diseño de un programa para discriminar litologías en un pozo usando redes neuronales*. Tesis en opción al grado de Ingeniero Geofísico, Universidad Simón Bolívar, Venezuela.

Trauth, M.H. MATLAB® Recipes for Earth Sciences. 2015. Institute for Earth and Environmental Science, Universidad de Postdam, Alemania,

Acerca de los autores:

Ernesto Alejandro Piloto-Zaldívar: Graduado de Ingeniero Geofísico en la Universidad Tecnológica de La Habana "José Antonio Echeverría" (2022). Actualmente trabajando en el Instituto de Geofísica y Astronomía (IGA) donde ejerce el puesto de Adiestrado. Ha participado en eventos científicos nacionales, proyectos y servicios del instituto, así como en cursos de Astronomía. Número ORCID: 0000-0003-0162-7529.

Omar González Trujillo: Graduado de Ingeniero Geofísico en la Universidad Tecnológica de Habana José Antonio Echeverría (2019). Actualmente trabajando en el Instituto de Geofísica y Astronomía (IGA) donde ejerció el puesto de Especialista en Ciencia, Tecnología y Medio Ambiente y posterior fue ascendido a Jefe del Departamento de Geofísica Espacial. Cuenta con 3 años de experiencia laboral. Además ha participado en eventos nacionales e internacionales relacionados con temas de su especialidad. Número ORCID: 0000-0001-8811-1394.

Oscar Alejandro Sánchez Sánchez: Graduado de Ingeniero Geofísico en la Universidad Tecnológica de La Habana "José Antonio Echeverría" (2022). Actualmente trabajando en el Instituto de Geofísica y Astronomía (IGA) donde ejerce el puesto de Adiestrado. Ha participado en eventos científicos nacionales, proyectos y servicios del instituto, así como en cursos de Astronomía. Número de ORCID: 0000-0002-3653-4368.

Zulia Bárbara Bárcena Fonseca: Graduada en Educación en la especialidad Economía (1993). Se incorporó a trabajar en el Instituto de Geofísica y Astronomía en el año 2003 y pasó a ser Subdirectora Cientiífica en el año 2014. Ha trabajado en diversos proyectos

científicos además de dirigir y supervisar muchos de ellos, tiene gran conocimiento en las áreas de las Geociencias, Astronomía y Geofísica Espacial. Ha participado en eventos nacionales e internacionales relacionados con temas de su especialidad y del instituto. Número de ORCID: 0000-0002-2029-2133.