Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
47(183):400-411, abril-junio de 2023. doi: https://doi.org/10.18257/raccefyn.1849

Artículo original

Predicción del ciclo solar 25 mediante modelos ARIMA y redes neuronales LSTM

Prediction of solar cycle 25 using ARIMA models and LSTM neural networks

Samuel Tomas[1], Oliver Saavedra[2], Israel Espinoza[1]

[1] Departamento de Ciencias Exactas, Universidad Privada Boliviana, Cochabamba, Bolivia
[2] Facultad de Ingeniería y Arquitectura, Universidad Privada Boliviana, Cochabamba, Bolivia

Resumen

Se realizó un estudio para predecir el número de manchas solares en el Ciclo Solar 25 mediante el uso de dos modelos: un modelo de redes neuronales recurrentes Long short-term memory (LSTM) y un modelo Autoregressive Integrated Moving Average (ARIMA). Los datos utilizados para entrenar los modelos fueron obtenidos del sitio web del Centro Mundial de Datos SILSO, del Real Observatorio de Bélgica en Bruselas, desde 1749 hasta 2018. Nuestro modelo LSTM demostró un rendimiento excepcional (RMSE=3,6) en comparación con el mejor modelo ARIMA (RMSE=32,6). Esto demostró que nuestro modelo LSTM es significativamente más preciso en términos de predicción, con una mejora del 89% en la reducción del RMSE. Según nuestro modelo LSTM, se prevé que el número máximo de manchas solares en el Ciclo Solar 25 ocurra en marzo de 2025, alcanzando un valor máximo de 182 manchas solares. En contraste, el modelo ARIMA predice que el máximo se alcanzará en diciembre de 2024, con un valor máximo de 99 manchas solares.

Palabras clave: Ciclo Solar 25; Manchas solares; Redes neuronales recurrentes; Predicción.

Abstract

A study was conducted to predict the number of sunspots in Solar Cycle 25 using two models: a Long short-term memory (LSTM) recurrent neural network model and an autoregressive integrated moving average (ARIMA) model. The data used to train the models was obtained from the website of the World Data Center SILSO, Royal Observatory of Belgium in Brussels, from 1749 to 2018. Our LSTM model demonstrated exceptional performance (RMSE=3.6) compared to the best ARIMA model (RMSE=32.6). This showed that our LSTM model is significantly more accurate in terms of prediction, with an 89% improvement in RMSE reduction. According to our LSTM model, the maximum number of sunspots in Solar Cycle 25 is expected to occur in March 2025, reaching a maximum value of 182 sunspots. In contrast, the ARIMA model predicts that the maximum will be reached in December 2024, with a maximum value of 99 sunspots.

Keywords: Solar Cycle 25; Sunspots; Recurrent neural networks; Prediction.

Introducción

La actividad solar no sólo afecta al clima espacial, sino que también tiene un impacto directo en la vida en la Tierra **Hathaway,** 2015; **Pulkkinen,** 2007. Las llamaradas solares y las eyecciones de masa coronal pueden causar daños en los sistemas de energía y comunicaciones **Gour *et al.*,** 2021; **Lybekk *et al.*,** 2012; **Solanki,** 2003; **Walterscheid,** 1989 como se ha visto en lo ocurrido en el año 1859, cuando el Sol estuvo muy activo y expulsó partículas muy cargadas que sobrecargaron las redes telegráficas, dejándolas inutilizables. Tal fue el impacto de estas partículas cargadas que se pudieron observar auroras boreales incluso en lugares poco habituales como Cuba, Madrid y Florida. Este evento es conocido en la comunidad de física solar como el evento Carrington, en honor al astrónomo inglés

Citación: Tomas S, Saavedra O, Espinoza I. Predicción del ciclo solar 25 mediante modelos ARIMA y redes neuronales LSTM. Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales. 47(183):400-411, abril-junio de 2023. doi: https://doi.org/10.18257/raccefyn.1849

Editor: Santiago Vargas Domínguez

***Correspondencia:**
Samuel Tomas; stomas@upb.edu

Recibido: 19 de diciembre de 2022
Aceptado: 3 de abril de 2023
Publicado en línea: 19 de abril de 2023

Richard Carrington, quien observó y registró el evento en detalle **Cliver & Dietrich,** 2013; **Moreno Cárdenas *et al.,*** 2016. Otro evento más reciente se registró en 1989, cuando una tormenta solar provocó un apagón en Quebec, y en 2003, cuando interrumpió las comunicaciones satelitales. Además, se ha descubierto que los ciclos solares también tienen efectos en la salud humana **Azcárate *et al.,*** 2016, como el aumento de la incidencia de enfermedades cardíacas y accidentes cerebrovasculares durante los máximos solares. Es esencial tener una predicción precisa de la actividad solar para tomar precauciones y minimizar los daños en las tecnologías y la salud humana **Juckett & Rosenberg**, 1993. Las implicaciones de la actividad solar son significativas y afectan a múltiples áreas de la vida en la Tierra, lo que subraya la importancia de seguir investigando en esta área.

La actividad solar se mide comúnmente a través del número de manchas solares **Usoskin**, 2017, lo cual ha sido registrado públicamente desde 1749. Para predecir el número de manchas solares, se han utilizado varias técnicas tanto métodos de aprendizaje no profundo como el modelo Autoregressive Integrated Moving Average ARIMA **Box & Jenkins**, 1976, como de aprendizaje profundo como Long Short-Term Memory LSTM **Hochreiter & Schmidhuber**, 1997, siendo estos últimos preferidos debido a su capacidad para extraer representaciones complejas de los datos. A pesar de que existen varios métodos para la predicción del número de manchas solares los modelos ARIMA y los métodos de aprendizaje profundo como LSTM son los más comúnmente utilizados debido a su capacidad para modelar patrones complejos en los datos **Han *et al.,*** 2019; **Li *et al.,*** 2021. Además, estos modelos han demostrado tener un alto grado de precisión en la predicción del número de manchas solares, lo que los hace valiosos para la toma de decisiones en diversas áreas que dependen de la actividad solar, como la energía y las comunicaciones.

Entre algunas investigaciones recientes en el campo de la predicción del número de manchas solares, podemos destacar el trabajo realizado por **Benson *et al*.,** 2020 en el que combinaron redes neuronales WaveNet y Long Short-Term Memory (LSTM) para predecir el número de manchas solares a partir de datos de series de tiempo desde 1749 hasta 2019. Los resultados de su modelo indican que el próximo Ciclo Solar 25 tendrá un número máximo de manchas solares cercano a 106 ± 19,75.

Por otro lado, **Prasad *et al.,*** 2022 utilizó el modelo de memoria a largo plazo basado en aprendizaje profundo LSTM para predecir la fuerza y el tiempo máximo del ciclo solar 25. Para ello, empleó datos mensuales del número de manchas solares. Su predicción indica que el Ciclo Solar 25 será más fuerte que el ciclo 24 y más débil que el ciclo 23, alcanzando su pico en agosto de 2023 ± 2 meses con una amplitud de número de manchas solares de 171, 9 ± 3, 4, lo que representa un aumento del 47% con respecto al ciclo 24.

Otra investigación interesante es la de **Dani & Sulistiani**, 2019, quienes utilizaron cuatro métodos de regresión de aprendizaje automático diferentes (Regresión Lineal, Bosque Aleatorio, Función de Base Radial y Máquina de Soporte Vectorial) para predecir el número de manchas solares. Los resultados de sus predicciones indican que el máximo del Ciclo Solar 25 se producirá en septiembre de 2023, con un número de manchas solares de 159, 4 ± 22, 3.

Finalmente, **Dang *et al***., 2022 y su equipo compararon tres modelos basados en aprendizaje no profundo, cuatro modelos populares de aprendizaje profundo y su modelos de ensamble para predecir el número de manchas solares. Su modelo XGBoost-DL logró el mejor rendimiento de predicción (RMSE = 25,70 y MAE = 19,82), superando al mejor modelo basado en aprendizaje no profundo SARIMA (RMSE = 54,11 y MAE = 45,51) y al mejor modelo de aprendizaje profundo Informer (RMSE = 29,90 y MAE = 22,35). Según este modelo, se espera que el Ciclo Solar 25 alcance su máximo con un número de manchas solares de 133,47 en mayo de 2025.

Si bien se han logrado avances en la predicción de manchas solares para el ciclo 25, todavía existen limitaciones significativas que afectan la calidad de las series históricas utilizadas y la precisión de los indicadores de error RMSE. Por esta razón, es necesario un enfoque más robusto y con datos más completos, utilizando modelos avanzados y técnicas sofisticadas para analizar dichos datos.

Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
47(183):400-411, abril-junio de 2023. doi: https://doi.org/10.18257/raccefyn.1849

En la predicción de series temporales, los modelos LSTM y ARIMA se encuentran entre las herramientas más prometedoras debido a su capacidad para mejorar significativamente la precisión. Los modelos LSTM son especialmente útiles para analizar patrones complejos y de largo plazo en los datos, mientras que los modelos ARIMA son ideales para analizar datos estacionarios y predecir valores futuros basados en patrones históricos.

En este estudio se ha llevado a cabo una comparación exhaustiva de ambos enfoques, LSTM y ARIMA. El objetivo principal es determinar cuál de los dos enfoques es más adecuado para predecir el número de manchas solares durante el ciclo 25.

Se espera que los resultados de este estudio proporcionen una comparación rigurosa de estos dos enfoques de modelado diferentes, lo que permitirá comprender mejor la variabilidad solar y sus efectos en nuestro clima espacial.

Métodos

En esta sección, se describirán los métodos utilizados para llevar a cabo la investigación sobre la predicción de manchas solares.

Modelo ARIMA

Una serie de tiempo es una secuencia de datos observados en momentos sucesivos en el tiempo, que se utilizan para analizar y predecir el comportamiento de un fenómeno en particular. En nuestro caso de estudio, la serie temporal corresponde al número de manchas solares observadas en el sol en períodos de 11 años, que se conocen como ciclos solares. El objetivo es utilizar los datos históricos para hacer una predicción del número de manchas solares en el ciclo solar 25.

El modelo ARIMA es un modelo econométrico utilizado para analizar y predecir series de tiempo **Shumway & Stoffer**, 2011. Se basa en la descomposición de una serie temporal en tres componentes: tendencia, estacionalidad y error. La tendencia representa la evolución a largo plazo de la serie, la estacionalidad describe patrones que se repiten a lo largo del tiempo y el error es la variación aleatoria no explicada por la tendencia y la estacionalidad.

El modelo ARIMA se construye a partir de la combinación de tres modelos básicos: el modelo AR (autorregresivo), el modelo de media móvil (MA) y el modelo de diferenciación (I). El modelo AR se basa en la regresión de la serie temporal en sí misma, es decir, se utiliza información pasada de la serie para predecir valores futuros. El modelo MA se basa en la regresión de la serie de errores, es decir, utiliza la información pasada de los residuos para predecir valores futuros. Finalmente, el modelo de diferenciación se utiliza para transformar una serie no estacionaria en una serie estacionaria. La idea es que, si una serie no es estacionaria, se puede aplicar una diferencia para eliminar la tendencia y la estacionalidad, dejando solo el error aleatorio.

Para aplicar el modelo ARIMA al ciclo solar 25, primero se realiza un análisis exploratorio de los datos históricos de las manchas solares. Se identifica si la serie temporal presenta tendencia y/o estacionalidad y se determina el número de diferencias necesarias para convertir la serie en estacionaria. Luego, se ajusta el modelo ARIMA a los datos estacionarios y se realiza una validación cruzada para evaluar la precisión del modelo en la predicción del número de manchas solares en el ciclo solar 25 **Montgomery *et al.,*** 2015.

Descripción de los datos

Se empleó un conjunto de datos recopilado a partir del número promedio mensual de manchas solares, obtenido del sitio web del Centro Mundial de Datos SILSO, del Real Observatorio de Bélgica en Bruselas, https://www.sidc.be/silso/ es importante destacar que el conjunto de datos es continuamente actualizado, sin embargo, para los propósitos de esta investigación, se consideró un conjunto de datos con 3240 registros que cubren el período desde enero de 1749 hasta diciembre de 2018.

Posteriormente, se dividió el conjunto de datos en dos grupos: entrenamiento y validación. El grupo de entrenamiento comprendió 2592 registros que cubren desde enero de 1749 hasta enero de 1965. El grupo de validación incluyó 648 registros, desde febrero de 1965 hasta diciembre de 2018.

Para realizar la predicción del ciclo solar 25, se utilizaron dos modelos: Redes neuronales recurrentes LSTM y modelos ARIMA. Se emplearon tanto métodos estadísticos como modelos de aprendizaje profundo para ajustar los datos y predecir el número de manchas solares en el futuro. El objetivo principal es identificar el modelo más preciso en términos de la métrica RMSE, con el fin de realizar pronósticos precisos del número de manchas solares durante el ciclo solar 25, desde enero de 2019 hasta diciembre de 2030. Este ciclo solar es de gran importancia para la comunidad científica, por lo que se busca obtener resultados confiables y precisos en la predicción de las manchas solares durante este período.

Implementación

Para iniciar la preparación de los datos y aplicar el modelo ARIMA, iniciamos con la visualización de la serie de tiempo del número de manchas solares mensuales desde el año 1749 hasta el año 2018, lo cual se puede observar en la **figura 1**. Este enfoque de análisis completo permite capturar tanto las tendencias históricas como las variaciones estacionales y cíclicas de la serie de tiempo, lo que es esencial para mejorar la precisión de las predicciones.

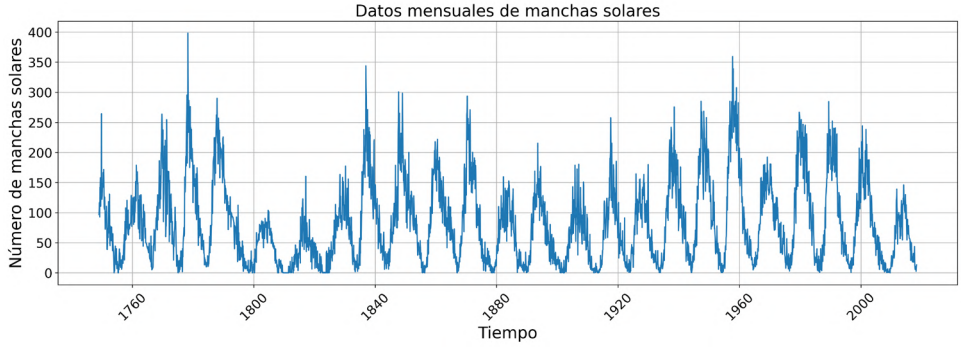

Figura 1. Recuento de manchas solares mensuales, registrado desde el año 1749 hasta el año 2018

Al analizar el gráfico del recuento de manchas solares, ver **figura 1**, se observa una periodicidad en los ciclos solares, con un periodo aproximado de 11 años. Cada ciclo solar tiene una fase de máxima actividad (con un mayor número de manchas solares) y una fase de mínima actividad (con un menor número de manchas solares).

Además, se ha observado una variabilidad en la duración e intensidad de estos ciclos solares a lo largo de la historia, lo que puede afectar la actividad y el clima espacial. Por ejemplo, el ciclo solar número 19 (entre 1954 y 1964) fue uno de los más intensos del siglo XX, con un máximo solar registrado en 1958.

Esta información está respaldada por estudios y observaciones realizadas por la comunidad científica, como el trabajo de **Hathaway *et al.*,** 1999, que estableció un modelo de predicción de la actividad solar basado en el análisis de los ciclos solares históricos y su duración e intensidad. Además, el estudio de **Eddy**, 1976 sugiere que el mínimo solar registrado entre 1645 y 1715 (conocido como el ”Mínimo de Maunder”) coincidió con una época de enfriamiento global conocida como la Pequeña Edad de Hielo.

A simple vista, la serie de tiempo del recuento de manchas solares no parece mostrar una tendencia clara a lo largo del tiempo, lo que sugiere que no sería necesario realizar

Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
47(183):400-411, abril-junio de 2023. doi: https://doi.org/10.18257/raccefyn.1849

una diferenciación para estabilizar la varianza. Sin embargo, para confirmar esto, se puede utilizar la prueba estadística de Dickey-Fuller (ADF), que se utiliza para evaluar la estacionariedad de la serie de tiempo.

En la prueba ADF, el valor p indica la probabilidad de que la serie de tiempo sea no estacionaria. Un valor de p menor a 0,05 se considera suficiente para rechazar la hipótesis nula de no estacionariedad. Para la serie de tiempo de manchas solares, se ha encontrado que el valor $p = 0, 02$ es menor que 0,05, lo que sugiere que la serie de tiempo es estacionaria y no requiere diferenciación.

Es importante tener en cuenta que la identificación de la estacionariedad es crucial para el modelado de series de tiempo, ya que los modelos ARIMA se basan en la suposición de que la serie de tiempo es estacionaria. Una serie no estacionaria puede llevar a modelos inapropiados y predicciones poco precisas **Box & Jenkins**, 1976; **Dickey & Fuller**, 1979.

Además de evaluar la estacionariedad de la serie de tiempo, también es importante considerar la homogeneidad de la varianza para asegurarnos de que los errores del modelo no presenten una varianza que cambie con el tiempo. En nuestro caso, hemos utilizado la prueba de homogeneidad de varianza de Levene para evaluar la homocedasticidad de la serie de tiempo del recuento de manchas solares. El resultado indica que la hipótesis nula de homogeneidad de varianza no se rechaza, ya que el valor de significancia obtenido es de $p > 0, 66$. Por lo tanto, podemos asumir que la serie de tiempo es homocedástica.

Una vez confirmada la estacionariedad y homogeneidad de varianza de la serie de tiempo de recuento de manchas solares, podemos proceder con la identificación del modelo. En particular, podemos afirmar que la serie tiene media constante y no presenta tendencia, lo que es consistente con la naturaleza cíclica de los datos. Además, como mencionamos previamente, se ha confirmado la homogeneidad de la varianza de la serie. De esta manera, podemos proceder a la identificación del modelo ARIMA que mejor se ajuste a la serie de tiempo.

Identificar el modelo adecuado para una serie de tiempo es un proceso crítico en el análisis de series de tiempo. Existen varias técnicas que se utilizan para la identificación del modelo ARIMA, como el análisis de la función de autocorrelación (ACF) y la función de autocorrelación parcial (PACF), entre otras. Estas técnicas permiten seleccionar los valores apropiados para los parámetros p, d, y q del modelo ARIMA **Box & Jenkins**, 1976.

Para llevar a cabo la identificación del modelo ARIMA más adecuado, dividimos los datos disponibles en dos conjuntos: uno que utilizaremos para el ajuste del modelo y otro que emplearemos para su validación. En concreto, utilizamos el 80% de los datos para el ajuste y el 20% restante para la validación.

Una vez definidos los conjuntos de entrenamiento y validación, procedemos a la identificación del modelo adecuado. Para ello, analizamos las gráficas de la función de autocorrelación (ACF) y de la función de autocorrelación parcial (PACF) ver **figura 2**, las cuales nos permiten determinar los valores de los parámetros p, d y q del modelo ARIMA. En función de las características de las gráficas, se pueden identificar distintos modelos que se ajusten a los datos **Abdel-Rahman & Marzouk**, 2018.

Es importante destacar que la elección del modelo final no solo se basa en el análisis de las gráficas **figura 2**, sino que también se tienen en cuenta otros criterios, como el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC), entre otros.

Durante el proceso de identificación del modelo, se llevó a cabo un análisis de las gráficas de autocorrelación y autocorrelación parcial, tal como se muestra en la **figura 2**. Además, se emplearon técnicas como AutoARIMA y Búsqueda en Cuadrícula para encontrar el modelo ARIMA óptimo. En esta última técnica, se estableció un rango de valores para los parámetros p, d y q mediante p = range(0, 5), d = range(0, 5) y q = range(0, 5), respectivamente. Se iteró a través de todas las combinaciones posibles dentro de estos rangos para encontrar el modelo con el valor más bajo de AIC. El número total de iteraciones realizadas fue de 125, lo que permitió la identificación de varios modelos potenciales. En la **tabla 1**, se presentan los modelos más prometedores.

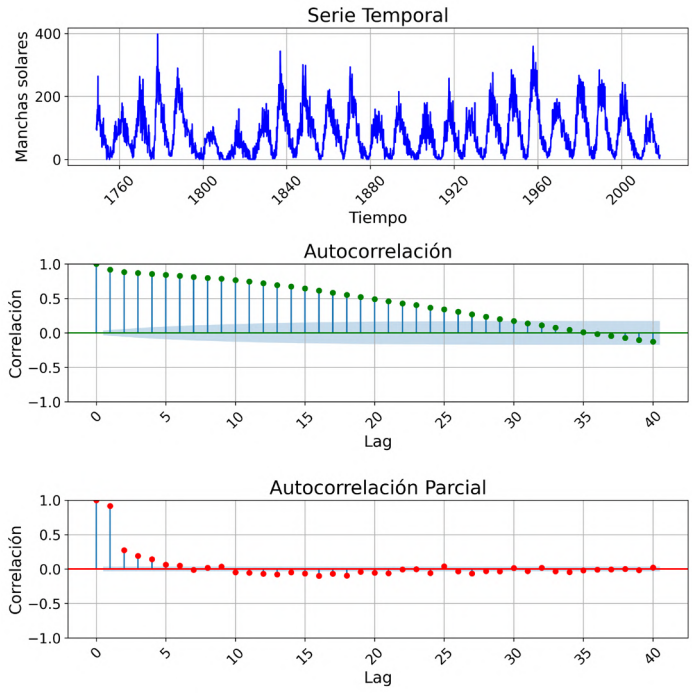

Figura 2. La autocorrelación simple y la autocorrelación parcial de la serie temporal muestran valores cercanos a 1 en los primeros intervalos de tiempo, lo que indica una fuerte correlación entre los valores pasados y presentes de la serie. Este patrón se puede observar fácilmente en el gráfico de autocorrelación, y se utilizó para identificar la presencia de patrones estacionales o de tendencias en la serie temporal, y la autocorrelación parcial se utilizó para eliminar el efecto de la correlación indirecta de los intervalos intermedios, permitiendo una mejor identificación de la verdadera relación entre los valores pasados y presentes de la serie

Tabla 1. Modelos candidatos y sus valores correspondientes de AIC, BIC y métrica RMSE

Modelo ARIMA	**Orden (p, d, q)**	**AIC**	**BIC**	RMSE
ARIMA 1	(3, 0, 2)	10071	10113	32,6
ARIMA 2	(1, 0, 8)	11024	11001	45,2
ARIMA 3	(2, 0, 1)	15071	15113	67

Luego de comparar los modelos candidatos, se seleccionó como modelo óptimo ARIMA (3, 0, 2), ya que presentaba el valor más bajo de AIC (10071) y de BIC (10113).

Para evaluar la calidad del ajuste del modelo, se realizó la prueba de Kolmogorov Smirnov, obteniéndose un valor de significancia de 0, 98. Este resultado indica que los residuos del modelo siguen una distribución normal, lo cual es un indicador importante de que el modelo ARIMA es adecuado para la serie de tiempo analizada.

Además de realizar el análisis de normalidad de los residuos, se compararon los modelos utilizando métricas de evaluación de ajuste. Se calculó el RMSE para cada modelo, lo que proporciona una medida de la precisión de las predicciones al cuantificar la diferencia entre las predicciones del modelo y los valores reales. Se encontró que el modelo ARIMA (3,0,2) presentó el valor más bajo de RMSE, lo que indica un mejor ajuste a los datos. Los valores de RMSE y AIC de los tres modelos más representativos se presentan en la **tabla 1**.

A continuación, se muestra una comparación entre las predicciones de tres modelos diferentes: ARIMA(2,0,1), ARIMA(1,0,8) y ARIMA(3,0,2) ver **figura 3**.

Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
47(183):400-411, abril-junio de 2023. doi: https://doi.org/10.18257/raccefyn.1849

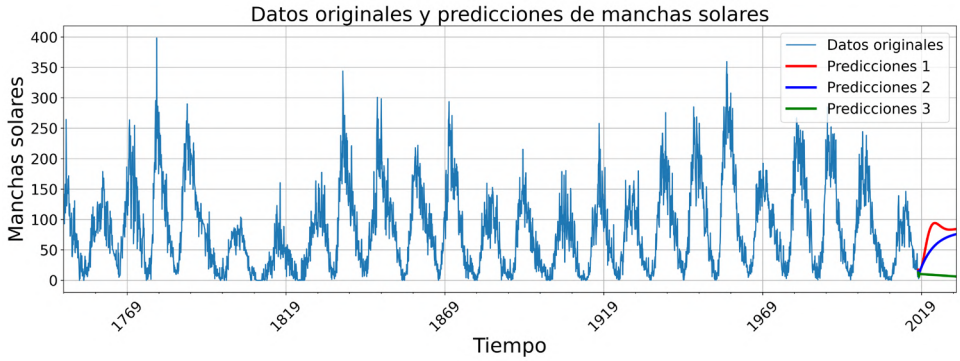

Figura 3. Predicción correspondiente a ARIMA(2,0,1) (predicción 3) no se ajusta bien a los datos, mientras que las predicciones correspondientes a ARIMA(1,0,8) (predicción 2) y ARIMA(3,0,2) (predicción 1) tienen un mejor ajuste, según sus valores de RMSE y AIC. Además, se puede visualizar un mejor ajuste en el gráfico para estas dos últimas predicciones.

Estos resultados indican que el modelo ARIMA (3,0,2) presenta el mejor ajuste en la serie temporal analizada.

Resultados del modelo ARIMA

La elección del modelo ARIMA (3, 0, 2) como el mejor modelo de pronóstico se basó en su desempeño sobresaliente en la métricas de evaluación, como el menor RMSE y los mejores valores de AIC y BIC en comparación con los otros modelos considerados. Además, esta elección se alinea con la metodología de **Box & Jenkins**, 1976. Por lo tanto, se presenta a continuación la **figura 4** que muestra tanto los valores observados de la serie como las predicciones generadas por el modelo ARIMA (3, 0, 2).

En la **figura 4** se muestra la curva estimada de los valores predichos para el número mensual de manchas solares durante el período de 2019-2030. Para obtener estas predicciones se utilizó el modelo ARIMA (3, 0, 2) ajustado a los datos mensuales del número de manchas solares desde 1749 hasta 2018. Según los resultados obtenidos, se espera un máximo de 99 manchas solares en diciembre de 2024. Es importante destacar que estas predicciones cuentan con un nivel de confianza del 95%, lo que indica que hay una buena probabilidad de que los resultados sean precisos y confiables.

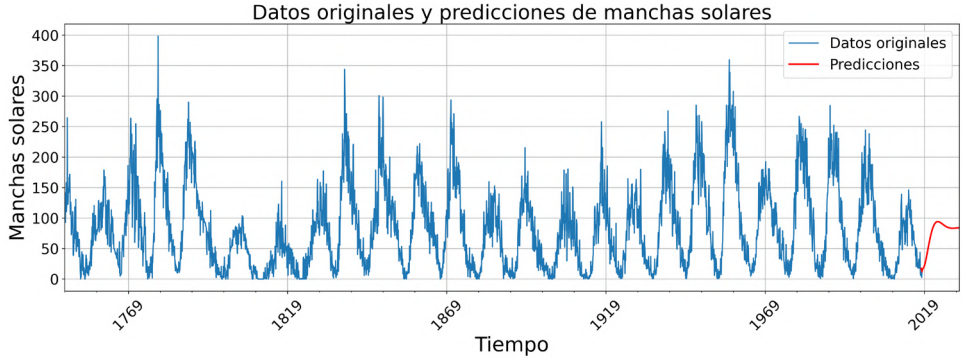

Figura 4. Predicción del número de manchas solares mensuales para el ciclo 25, obtenida mediante modelos ARIMA aplicados al período 2019-2030

Redes neuronales recurrentes LSTM

Las redes neuronales recurrentes (RNNs) son un tipo de modelo de aprendizaje profundo que se utiliza para analizar datos secuenciales, como el procesamiento del lenguaje natural o la predicción de series de tiempo. Sin embargo, la capacidad de las RNNs para manejar datos a largo plazo se ve comprometida por el problema del gradiente que desaparece, que ocurre cuando el gradiente se hace cada vez más pequeño a medida que se propaga hacia atrás en la red.

Para superar este problema, **Hochreiter & Schmidhuber**, 1997 propusieron la memoria a corto y largo plazo (LSTM), una variante de las RNNs que ha demostrado ser efectiva para el procesamiento de datos secuenciales a largo plazo. La arquitectura de LSTM se compone de celdas que contienen unidades de procesamiento y una memoria a largo plazo que puede ser controlada por tres puertas diferentes: entrada, salida y olvido **Goodfellow *et al.,*** 2016.

El modelo matemático de LSTM se define como una función no lineal que transforma la entrada actual, el estado anterior y la memoria a largo plazo en una salida y un estado actualizado **Hochreiter & Schmidhuber**, 1997. El cálculo de esta función implica la operación de multiplicación de matrices y la aplicación de funciones de activación, como la función sigmoide o la tangente hiperbólica.

Para el desarrollo del estudio, se emplearon redes neuronales LSTM. El procesamiento de los datos y el entrenamiento del modelo se realizaron utilizando diversas herramientas y bibliotecas en Python.

Se utilizó una unidad de procesamiento gráfico (GPU) Nvidia GeForce GTX 1080Ti, y se aprovechó la potencia de procesamiento de la GPU mediante la biblioteca de programación paralela CUDA de Nvidia. Asimismo, se utilizó la biblioteca de redes neuronales CudNN de Nvidia para acelerar el procesamiento de redes neuronales en la GPU, permitiendo así el procesamiento rápido y eficiente de los datos y el entrenamiento del modelo LSTM.

Para la visualización de los datos y los resultados de las predicciones, se importaron las bibliotecas NumPy, Pandas, TensorFlow, Keras y matplotlib.pyplot. El paquete pandas-datareader se empleó para descargar y cargar los datos de manchas solares, mientras que Pandas se utilizó para la limpieza y preparación de los mismos para su uso en el modelo LSTM. Por otro lado, Numpy se empleó para la manipulación de matrices y operaciones matemáticas, así como para la preparación de datos y la realización de cálculos en el modelo LSTM.

TensorFlow se utilizó para construir y entrenar modelos de aprendizaje profundo, haciendo uso de la GPU y las bibliotecas CUDA y CudNN para acelerar el entrenamiento del modelo. Para la construcción de modelos de redes neuronales en Keras, se empleó la sub-biblioteca tensorflow.keras.models, mientras que la subbiblioteca tensorflow.keras.layers se utilizó para definir las diferentes capas de redes neuronales en el modelo LSTM. Para la búsqueda de hiperparámetros y la optimización de modelos, se utilizó la herramienta Keras-tuner. Además, se empleó el módulo math de Python para funciones matemáticas básicas, y la biblioteca sklearn.metrics para evaluar el rendimiento del modelo LSTM en la predicción de manchas solares.

Descripción de los datos

En el preprocesamiento de los datos de las manchas solares, utilizaremos datos históricos del ciclo solar y diseñaremos una red neuronal recurrente LSTM capaz de aprender y modelar las relaciones temporales en los datos.

Para asegurar que la red neuronal aprenda de manera eficiente, normalizamos los datos utilizando la técnica de normalización Min-Max, la cual se define mediante la fórmula

$$X_{std} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

Este proceso escala los datos a un rango de valores entre 0 y 1, como se muestra en la **figura 5**, y es fundamental para llevar todas las entradas de la red a magnitudes similares y mejorar el desempeño de la red neuronal.

Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
47(183):400-411, abril-junio de 2023. doi: https://doi.org/10.18257/raccefyn.1849

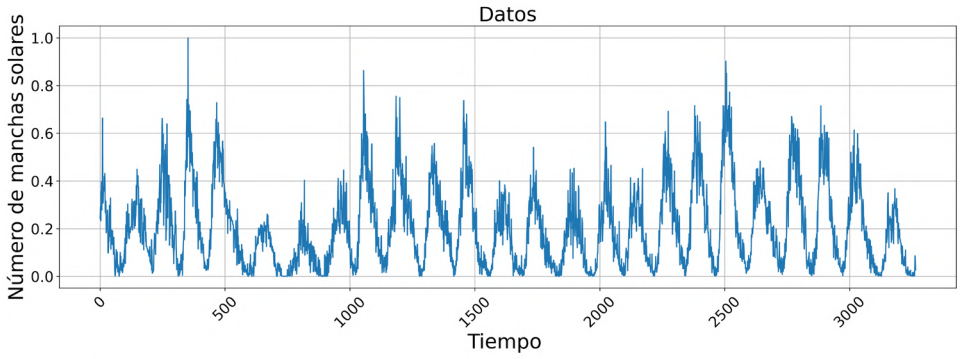

Figura 5. Serie de datos normalizados del número de manchas solares registradas en el sol desde 1749 hasta 2021

Para la construcción de la red neuronal LSTM, se dividió el conjunto de datos de manchas solares en dos grupos: el conjunto de entrenamiento y el conjunto de validación. El conjunto de entrenamiento se seleccionó de manera consecutiva y ordenada, representando el 80% de la serie temporal, y se utilizó para el diseño y construcción de la red neuronal.

Para evaluar la capacidad de generalización de la red, se reservó el 20% restante de los datos, constituyendo el conjunto de validación. Este conjunto de datos se utilizó para evaluar el desempeño de la red neuronal en la detección de patrones en la serie temporal que no se encontraban en el conjunto de entrenamiento **Benson *et al.,*** 2020.

Para implementar nuestro modelo de predicción del ciclo solar 25, hemos utilizado la biblioteca Keras en Python ”Keras, the Python deep learning API” https://keras.io/api/. En particular, hemos utilizado la función Sequential de Keras para construir nuestro modelo de red neuronal, y hemos incorporado capas LSTM para capturar las dependencias temporales en los datos de la serie de tiempo **Hochreiter & Schmidhuber**, 1997. Además, para convertir el problema de serie de tiempo en una tarea de regresión, hemos utilizado variables de rezago y sin rezago en nuestro modelo **Lütkepohl**, 2005. Al tomar en cuenta las variables de rezago y sin rezago, hemos transformado el problema de serie de tiempo en una tarea de regresión y ajustado los pesos de la red neuronal para predecir el número de manchas solares en el futuro. El entrenamiento de un modelo LSTM implica ajustar los parámetros del modelo utilizando datos de entrenamiento, con el objetivo de que el modelo pueda hacer predicciones precisas en nuevos datos. En este proceso, se utilizan diferentes arquitecturas de capas LSTM y una capa densa con diferentes números de neuronas. La optimización de los parámetros se realiza mediante el método de optimización Adam, que es un algoritmo de descenso de gradiente estocástico que actualiza los pesos de las capas de la red neuronal. La función de pérdida utilizada es el RMSE **Goodfellow *et al.,*** 2016. Se utiliza el buscador automático de hiperparámetros Tuner para seleccionar los mejores hiperparámetros del modelo. Durante el proceso de entrenamiento, se evalúa la pérdida en el conjunto de prueba para comparar el desempeño de cada modelo. En nuestro caso, se determinó que el número de épocas para obtener los mejores resultados fue de 500 y el tamaño del lote fue de 120. Ademas el entrenamiento del modelo LSTM ha mostrado una mejora constante en la capacidad predictiva a medida que se han ajustado los parámetros durante las diferentes épocas de entrenamiento.

Para evaluar el desempeño del modelo, se utilizaron los datos de validación que no se utilizaron en el entrenamiento. El modelo alcanzó una precisión del 90, 5% en la predicción del ciclo solar 25.

Resultados del modelo LSTM

En esta sección se llevó a cabo la predicción del ciclo solar 25 utilizando el modelo de red neuronal LSTM previamente entrenado con los datos normalizados de manchas solares

desde 1749 hasta diciembre de 2018. La selección de hiperparámetros y la validación del modelo propuesto fueron realizadas previamente. Se predijo el comportamiento del ciclo solar 25 para los próximos 11 años, es decir, hasta diciembre de 2030.

En la **figura 6** se puede apreciar claramente la predicción realizada por el modelo.

La predicción del modelo LSTM sugiere que el ciclo solar 25 alcanzará su máximo de 182 manchas solares en marzo de 2025, lo que proporciona información valiosa para la investigación del clima espacial y la predicción de eventos solares.

Además, al comparar los resultados con el mejor modelo ARIMA con (RMSE=32,6), se encontró que nuestro modelo LSTM tuvo un rendimiento excepcional con un (RMSE=3,6) lo que indica una mejora significativa del 89% en la reducción del RMSE en la predicción. Por lo tanto, se puede afirmar que el modelo LSTM es significativamente más preciso en términos de predicción y puede ser una herramienta útil para pronosticar eventos en series de tiempo similares.

Según la **tabla 2**, se puede concluir que el modelo LSTM ha demostrado ser una técnica de predicción efectiva para las manchas solares. El RMSE de nuestro modelo LSTM fue de 3,6, que es significativamente más bajo que el mejor modelo ARIMA con un RMSE de 32,6. Además, los resultados de la predicción del modelo LSTM son comparables a los de otros autores en términos de RMSE y predicción. Por lo tanto, se puede afirmar que el modelo LSTM es una herramienta útil para pronosticar eventos en series de tiempo similares y es especialmente adecuado para la predicción del ciclo solar 25.

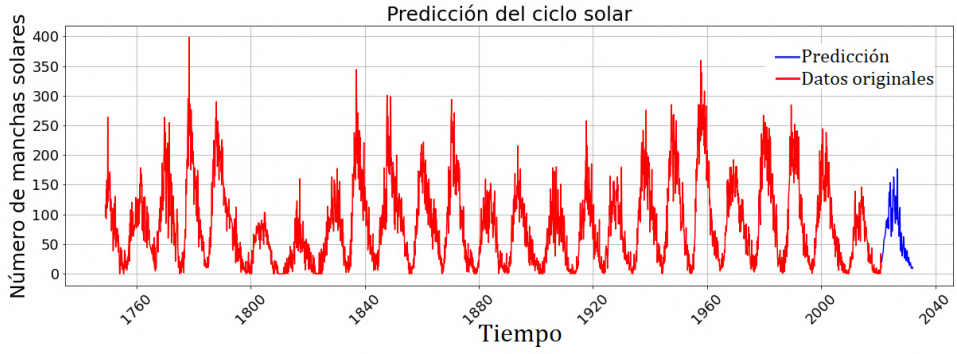

Figura 6. Predicción del ciclo solar 25, basada en datos desde 1749 hasta diciembre de 2018. Se ha pronosticado un período de 11 años hacia el futuro, hasta diciembre de 2030, y se estima que el máximo número de manchas solares será de 182 en marzo de 2025

Tabla 2. Predicciones del ciclo solar 25 utilizando modelos de aprendizaje profundo de diferentes autores

Modelo	RMSE	Predicción	Referencia
WaveNet + LSTM	4,42	106 ± 19,75	**Benson *et al.*,** 2020
LSTM	3,4	171,9 ± 3,4	**Prasad *et al.*,** 2022
LSTM	35,9	—	**Pala & Atici**, 2019
LSTM	6,12-2,45	114,3	**Wang *et al.*,** 2021
LSTM XGBoost-DL	25,7	133,47	**Dang *et al.*,** 2022

Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
47(183):400-411, abril-junio de 2023. doi: https://doi.org/10.18257/raccefyn.1849

Conclusiones

El ciclo solar, caracterizado por la variación en el número de manchas solares, es un fenómeno que ha sido objeto de estudio durante siglos debido a su influencia en nuestro planeta. La capacidad de predecir el número de manchas solares es importante no solo para el estudio del Sol, sino también para las aplicaciones en la industria eléctrica, satelital y de comunicaciones.

Se llevó a cabo un estudio para predecir el número de manchas solares en el Ciclo Solar 25 utilizando dos modelos: un modelo de redes neuronales recurrentes LSTM y el modelo ARIMA. Los resultados mostraron que el modelo LSTM superó significativamente al modelo ARIMA en términos de precisión de predicción, con una mejora del 89% en la reducción del RMSE. El modelo LSTM demostró un rendimiento excepcional, con un RMSE de 3,6, mientras que el modelo ARIMA obtuvo un RMSE de 32,6.

La capacidad predictiva de los modelos se evaluó utilizando la métrica RMSE. Se observó que el modelo de aprendizaje profundo LSTM no tiende a alejarse del valor real a medida que se prolonga en el tiempo, mientras que el modelo estadístico ARIMA muestra una tendencia a alejarse de la serie original a medida que se prolonga en el tiempo, lo que sugiere que es menos adecuado para el pronóstico de la serie.

Finalmente, según el modelo LSTM, se prevé que el número máximo de manchas solares en el Ciclo Solar 25 ocurra en marzo de 2025, alcanzando un valor máximo de 182 manchas solares. En contraste, el modelo ARIMA predice que el máximo se alcanzará en diciembre de 2024, con un valor máximo de 99 manchas solares, este resultado tiene importantes implicaciones para la industria eléctrica, satelital y de comunicaciones que dependen de la actividad solar.

Agradecimientos

Queremos expresar nuestro sincero agradecimiento a los revisores anónimos por el tiempo, esfuerzo y conocimiento que han dedicado a la revisión del manuscrito. Sus comentarios y sugerencias han sido extremadamente útiles para mejorar la calidad del trabajo.
Tambien expresar el agradecimiento a Gonzalo Vargas por mostrarnos el camino en el apasionante mundo de las manchas solares.

Contribución de los autores

ST, OS e IE planearon y diseñaron el estudio y analizaron la información. Todos los autores participaron en el trabajo de elaboración, la escritura del manuscrito, su revisión y la aprobación de la versión final.

Conflicto de intereses

Los autores declaran no tener conflictos de interés.

Referencias

Abdel-Rahman, H. I., Marzouk, B. A. (2018). Statistical method to predict the sunspots number. *NRIAG Journal of Astronomy and Geophysics*, *7*(2), 175-179. https://doi.org/10.1016/j.nrjag.2018.08.001

Azcárate, T., Mendoza, B., Levi, J. (2016). Influence of geomagnetic activity and atmospheric pressure on human arterial pressure during the solar cycle 24. *Advances in Space Research*, *58*(10), 2116-2125.

Benson, B., Pan, W. D., Prasad, A., Gary, G. A., Hu, Q. (2020). Forecasting Solar Cycle 25 Using Deep Neural Networks. *Solar Physics*, *295*(5), 65. https://doi.org/10. 1007/s11207-020-01634-y

Box, G. E. P., Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. Retrieved February 28, 2023, from https://dialnet.unirioja.es/servlet/libro?codigo=375102

Cliver, E., Dietrich, W. (2013). The 1859 space weather event revisited: Limits of extreme activity. *Journal of Space Weather and Space Climate*, *3*, 31–. https://doi.org/10. 1051/swsc/2013053

Dang, Y., Chen, Z., Li, H., Shu, H. (2022). A comparative study of non-deep learning, deep learning, and ensemble learning methods for sunspot number prediction. *Applied Artificial Intelligence*, *36*(1), 2074129.

Dani, T., Sulistiani, S. (2019). Prediction of maximum amplitude of solar cycle 25 using machine learning. *Journal of Physics: Conference Series*, *1231*(1), 012022.

Dickey, D. A., Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427-431. https://doi.org/10.1080/01621459.1979.10482531

Eddy, J. A. (1976). The maunder minimum. *Science*, *192*(4245), 1189-1202. https://www.science.org/doi/10.1126/science.192.4245.1189

Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep learning*. MIT press.

Gour, P. S., Singh, N. P., Soni, S., Saini, S. M. (2021). Observation of coronal mass ejections in association with sun spot number and solar flares. *IOP Conference Series: Materials Science and Engineering*, *1120*(1), 012020.

Han, Z., Zhao, J., Leung, H., Ma, K. F., Wang, W. (2019). A review of deep learning models for time series prediction. *IEEE Sensors Journal*, *21*(6), 7833-7848.

Hathaway, D. H. (2015). The Solar Cycle [arXiv:1502.07020 [astro-ph]]. *Living Reviews in Solar Physics*, *12*(1), 4. https://doi.org/10.1007/lrsp-2015-4

Hathaway, D. H., Wilson, R. M., Reichmann, E. J. (1999). A synthesis of solar cycle prediction techniques. *Journal of Geophysical Research: Space Physics*, *104*, 22375-22388. https://doi.org/10.1029/1999JA900313

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*, 1735-80. https://doi.org/10.1162/neco.1997.9.8.1735

Juckett, D. A., Rosenberg, B. (1993). Correlation of human longevity oscillations with sunspot cycles. *Radiation Research*, *133*(3), 312-320.

Li, Q., Wan, M., Zeng, S.-G., Zheng, S., Deng, L.-H. (2021). Predicting the 25th solar cycle using deep learning methods based on sunspot area data. *Research in Astronomy and Astrophysics*, *21*(7), 184.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer. https://doi.org/10.1007/978-3-540-27752-1

Lybekk, B., Pedersen, A., Haaland, S., Svenes, K., Fazakerley, A. N., Masson, A., Taylor, M., Trotignon, J.-G. (2012). Solar cycle variations of the cluster spacecraft potential and its use for electron density estimations. *Journal of Geophysical Research: Space Physics*, *117*(A1).

Montgomery, D. C., Jennings, C. L., Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley & Sons.

Moreno Cárdenas, F., Cristancho Sánchez, S., Vargas Domínguez, S. (2016). The grand aurorae borealis seen in Colombia in 1859. *Advances in Space Research*, *57*(1), 257-267. https://doi.org/10.1016/j.asr.2015.08.026

Pala, Z., Atici, R. (2019). Forecasting Sunspot Time Series Using Deep Learning Methods. *Solar Physics*, *294*(5), 50. https://doi.org/10.1007/s11207-019-1434-6

Prasad, A., Roy, S., Sarkar, A., Panja, S. C., Patra, S. N. (2022). Prediction of solar cycle 25 using deep learning based long short-term memory forecasting technique. *Advances in Space Research*, *69*(1), 798-813.

Pulkkinen, T. (2007). Space Weather: Terrestrial Perspective. *Living Reviews in Solar Physics*, *4*(1), 1. https://doi.org/10.12942/lrsp-2007-1

Shumway, R., Stoffer, D. (2011). Arima models', time series analysis and its applications.

Solanki, S. K. (2003). Sunspots: An overview. *Astronomy & Astrophysics Review*, *11*.

Usoskin, I. (2017). A history of solar activity over millennia. *Living Review in Solar Physics*, *14*, 3.

Walterscheid, R. (1989). Solar cycle effects on the upper atmosphere-implications for satellite drag. *Journal of spacecraft and rockets*, *26*(6), 439-444.

Wang, Q.-J., Li, J.-C., Guo, L.-Q. (2021). Solar cycle prediction using a long short-term memory deep learning model [Publisher: National Astronomical Observatories, CAS and IOP Publishing Ltd.]. *Research in Astronomy and Astrophysics*, *21*(1), 012. https://doi.org/10.1088/1674-4527/21/1/12