MODELOS DE MACHINE LEARNING PARA ESTIMAR LA RADIACIÓN SOLAR HORIZONTAL EN LA PAMPA HÚMEDA CON INFORMACIÓN SATELITAL MULTIESCALA

Paula Iturbide[1], Ximena Orsi[1], María José Denegri[1,2], Santiago Fioretti[1], Pablo Ruiz[1], Sergio Luza[1], Valeria Stern[1], Rodrigo Alonso-Suárez[3], Franco Ronchetti[4,5]

[1] Grupo de Estudios de la Radiación Solar (GERSolar), Instituto de Ecología y Desarrollo Sustentable (INEDES). Univ. Nacional de Luján, CP 6700, Buenos Aires, Argentina.

[2] Departamento de Tecnología, Universidad Nacional de Luján, Buenos Aires, Argentina.

[3] Laboratorio de Energía Solar, Dpto. de Física del CENUR Litoral Norte, Udelar, Uruguay.

[4] Instituto de Investigación en Informática LIDI, Universidad Nacional de La Plata, Buenos Aires, Argentina.

[5] Comisión de Investigaciones Científicas de la Pcia. de Buenos Aires (CICPBA), Buenos Aires, Argentina.

e-mail: paula.itur@gmail.com

RESUMEN: La falta de precisión en los datos de radiación solar puede tener un gran impacto en la rentabilidad de los proyectos de energía solar. Las redes de medición terrestre ofrecen información limitada por su distribución esparza en el territorio. Esto lleva a desarrollar modelos de estimación por imágenes satelitales, los cuales resuelven la espacialidad si son ajustados a mediciones terrestres de calidad. En este estudio, se desarrollan y validan modelos empíricos de aprendizaje automático para la estimación por satélite de radiación solar global horizontal, demostrando su utilidad y precisión en la región analizada. Estos modelos se alimentan con variables provenientes de imágenes satelitales GOES-16 y variables geométricas. Los resultados sugieren que, para ciertas combinaciones de variables satelitales de entrada, la información geométrica puede ser utilizada en forma implícita para realizar estimaciones precisas de la radiación solar. Debido al volumen de la información satelital disponible, desarrollamos un análisis de componentes principales para reducir la dimensionalidad. Para comparar el modelo propuesto adaptamos localmente las estimaciones del Heliosat-4 y del CIM-ESRA al sitio, y también implementamos el modelo CIM-McClear. Los resultados muestran una superioridad de desempeño del modelo de aprendizaje automático propuesto, demostrando que es capaz de extraer información de la multiescala espacial satelital. Por otro lado, la mejora de desempeño obtenida es leve, lo que muestra la dificultad en seguir mejorando el desempeño de la estimación satelital de radiación solar.

Palabras clave: Radiación solar, aprendizaje automático, Imágenes satelitales. GOES16, GHI.

INTRODUCCIÓN

En el ámbito de la estimación de radiación solar por satélite coexisten tres enfoques: el físico, el estadístico y el híbrido. Los modelos físicos resuelven las ecuaciones de transferencia radiante en la atmósfera, utilizando información sobre los componentes atmosféricos que interactúan con la radiación solar (Perez R. et al., 2013). Los modelos estadísticos se basan en una serie de coeficientes ajustados empíricamente a datos terrestres tomando como entrada la información satelital. Por último, en los modelos híbridos o semi-empíricos la formulación del modelo tiene una base física, pero dependen de una serie de parámetros ajustables.

En la región de la Pampa Húmeda se han desarrollado modelos satelitales híbridos específicamente ajustados, como el CIM-ESRA y CIM-McClear (Laguarda et al., 2020). Otro modelo relevante en la región es el modelo físico Heliosat-4 (Qu et al., 2017), que ha sido extensamente evaluado (Gonzalez et

al., 2019; Laguarda et al., 2020, 2021; Sarazola et al., 2023). Los modelos CIMs, al utilizar imágenes GOES16 y estar ajustados localmente a la región, han sido evaluados con significativo mejor desempeño en la Pampa Húmeda que el Heliosat-4, que utiliza imágenes del satélite europeo Meteosat, obteniendo desvíos cuadráticos medios entre 16-17% (relativo a la media de las medidas) para estimaciones de irradiancia solar global horizontal (GHI) a escala 10-minutal. Reducir la incertidumbre por debajo de este límite ha demostrado ser un desafío.

Una posible forma de reducir esta incertidumbre es a través del uso de información satelital multi-escala espacial y algoritmos de aprendizaje automático como redes neuronales artificiales (RN), k vecinos más cercanos (kNN), vectores soporte de regresión (SVR), máquinas de aprendizaje extremas (ELM) y ensambles de árboles como random forest (RF) y gradient boosting (GB). En la actualidad es común en varias áreas el uso de estos algoritmos. En particular, para la estimación de la radiación solar se han utilizado principalmente con base en mediciones terrestres de otras variables meteorológicas como presión, temperatura de aire ambiente, heliofanía, humedad, precipitación, nubosidad vista desde tierra, velocidad del viento y/o evaporación, etc., y variables auxiliares como el día del año, latitud, longitud, altitud, modelos de cielo claro, y/o variables geométricas, entre otras (Raichijk, 2008; Sayago et al., 2011; Jiménez et al., 2017; Olivera et al., 2020). Las investigaciones que emplean estas técnicas con información satelital de nubosidad son escasas (Verbois et al., 2023).

Este trabajo es la continuación del artículo de Iturbide et al. (2023), donde se implementaron los algoritmos RN, RF y regresión lineal simple para la estimación de la GHI por satélite. Se utilizaron 38 variables de entrada de las cuales 19 estaban relacionadas con el factor de reflectancia y 19 con la reflectancia planetaria. Estas variables abarcaban diversas resoluciones espaciales, variando entre 0,01 y 0,9 grados de latitud y longitud. Además, se incluyeron el coseno del ángulo cenital y el modelo de cielo claro McClear como parte de las variables de entrada. Los resultados de dicho artículo mostraron un rendimiento superior por parte de la RN, seguida por el enfoque de RF, incluso después de la eliminación de variables como el modelo de cielo claro y el coseno del ángulo cenital. En contraste, la regresión lineal simple demostró un desempeño insuficiente al excluir estas variables, ya que carecía de la capacidad para reconstruir la referencia temporal esencial para la estimación de la GHI. Cabe mencionar que la comparación con modelos preexistentes en Iturbide et al. (2023) se llevó a cabo sin realizar la adaptación al sitio.

El objetivo de este artículo es mejorar el desempeño de los modelos de aprendizaje automático mediante la incorporación de variables de entrada previamente no consideradas, tales como el índice de nubosidad y un cálculo mejorado de la reflectancia planetaria. Además, se introduce un análisis de componentes principales que contribuye a reducir la dimensionalidad del conjunto de entrada, mejorando la eficacia de los modelos y condensando la información satelital en un conjunto reducido de variables. También se agrega el modelo GB para comparar su desempeño. Se procede a comparar el modelo resultante con los disponibles para la región, adaptados al sitio, y se incluye la implementación con ajuste local a medidas del modelo CIM-McClear. Para ello, se usa el mismo esquema de ajuste y testeo que el utilizado para los modelos de aprendizaje automático, de modo de realizar una comparativa justa.

METODOLOGÍA

Medidas en tierra y pre procesamiento de datos

En este estudio se utilizaron datos de la estación Luján (de la red GERSolar) correspondientes al periodo 2019-2021, adquiridos con piranómetro de la firma Kipp & Zonen modelo CMP21 - equipo de Clase A según la norma de clasificación de equipamiento para la medida de radiación solar (ISO 9060:2018)-, y un adquisidor de datos Campbell Scientific modelo CR1000. Se adoptó la escala temporal de 10 minutos, que es la cadencia temporal de las imágenes capturadas por el satélite GOES-16. Las integrales 10 minutales (en W/m^2) fueron sometidas a un algoritmo de control de calidad que consta de los cuatro filtros secuenciales mostrados en la Tabla 1, seguido de una revisión visual de las series para descartar períodos afectados por sombras u otros fallos. En Iturbide et al. (2023) se detalla lo que impone cada filtro. En la Tabla 2 se muestran los resultados del filtrado para la estación Luján, donde se indica además su ubicación precisa.

Tabla 1: Filtros aplicados a las medidas en tierra

Filtro	Criterio	Descripción
1	$\alpha_s > 7^{\circ}$	Mínima altura solar
2	$-2\mathrm{W/m^2} < \mathrm{I_h} < \mathrm{I_0}.1{,}2.\cos\theta_z^{1.2} + 50\mathrm{W/m^2}$	Cotas de la BSRN (Long y Shi, 2008)
3	$0\mathrm{W/m^2} < \mathrm{I_h} < \mathrm{I_h^{ESRA}}(\mathrm{TL} = 1{,}8)$	Cotas de un modelo de cielo claro
4	$\mathrm{ktp} < 0{,}89$	Cota del índice de claridad de Pérez

Tabla 2: Ubicación de la estación de medida y sus equipos de medición. Los N totales (luego de los filtros) corresponden a medidas integradas 10-minutales y el período corresponde a 2019-2021

Estación	Latitud (grados)	Longitud (grados)	Equipo	N Total
Luján, ARG	-34,558	-59,062	CMP21	62.592

Información satelital

Se utilizan las imágenes del canal visible (C02, centrado en 0,64 µm) del satélite meteorológico geoestacionario GOES-16. Este satélite forma parte de la red de satélites geoestacionarios para la observación de la Tierra que cubre todo el globo terráqueo y es administrado por la National Oceanic and Atmospheric Administration (NOAA) de los Estados Unidos. Desde el año 2018 este satélite genera imágenes para todo el continente americano con una cadencia temporal regular de entre 10 y 15 minutos. Se encuentra ubicado sobre el ecuador terrestre en la longitud -75°W. Su resolución espacial es variable a lo largo de la imagen, siendo de 500 m en su nadir. Sobre la región de la Pampa Húmeda el tamaño del píxel varía entre 1 y 3 km. El canal visible es el adecuado para la estimación de radiación solar debido a que la nubosidad diurna es claramente reconocible y cuantificable. Las nubes son típicamente más reflectivas que el fondo (la superficie terrestre), y por tanto, distinguibles.

Las dos variables típicas que se calculan a partir del canal visible de una imagen satelital son el factor de reflectancia (FR) y la reflectancia planetaria (RP). Esta última cantidad es también conocida como Albedo terrestre. El FR es una normalización de la radiancia medida por el satélite proveniente de cada píxel respecto al máximo que es capaz de medir (es decir, la radiación solar que incide sobre el tope de la atmósfera normalizada por la respuesta espectral del radiómetro en órbita). Se encuentra por tanto en el intervalo [0, 1] y contiene, además de información sobre nubosidad, información espacial sobre la iluminación variable del Sol sobre la Tierra. La cantidad RP contiene la normalización necesaria para eliminar esta dependencia espacial y es efectivamente la reflectividad de la Tierra, en su sentido físico estricto. Esta normalización se obtiene dividiendo a FR por el coseno del ángulo cenital solar.

La reflectancia planetaria es también normalizada para obtener el índice de nubosidad N de la siguiente manera:

$$N = (R - R_0)/(R_{max} - R_0) \tag{1}$$

donde R_0 es la reflectancia planetaria de fondo asociado a condiciones de cielo claro para cada celda y el parámetro R_{max} se asocia a condiciones de nubosidad total. Se utiliza aquí un modelo de fondo para el cálculo de R_0 específicamente ajustado al pixel objetivo (Alonso-Suárez et al., 2011) y un valor fijo de 0,8 para R_{max}, que fue optimizado para la estimación de GHI en la región (Laguarda et al., 2018).

Estas variables permiten caracterizar la nubosidad a partir de las imágenes satelitales y son las que se usan para estimar la radiación solar en toda condición de cielo.

Modelos estadísticos con aprendizaje automático

Definición de los conjuntos de entrenamiento y testeo. El propósito principal es que los algoritmos adquieran la capacidad de estimar la GHI mediante el ajuste a mediciones terrestres. Dado el carácter estacional anual de la GHI, se empleó un conjunto de datos abarcando dos de los tres años disponibles para el entrenamiento, mientras que el tercer año se reservó para el testeo. Se aplicaron todas las posibles combinaciones de años para evitar tres posibles sesgos. En primer lugar, se evitó un sesgo relacionado con la distribución aleatoria de los datos en los conjuntos de entrenamiento y testeo, ya que los datos de momentos consecutivos podrían presentar similitudes. En segundo lugar, se mitigó un sesgo vinculado a las particularidades de cada año, pues un año podría diferir significativamente de otro en términos de radiación solar. En tercer lugar, se abordó un sesgo relacionado con los datos faltantes: el año 2020 registró un hueco de dos meses, y para los años 2019 y 2021 se carece de información correspondiente al mes de diciembre.

Variables de entrada. Las variables de entrada utilizadas en los modelos de aprendizaje automático son el coseno del ángulo cenital (cosz), el factor de reflectancia (FR), la reflectancia planetaria y el índice de nubosidad. Se trabaja con dos índices de nubosidad N1 y N2 y dos factores de reflectancia R y RC. Las variables cosz, FR y R son las mismas que se utilizaron en el artículo de Iturbide et al. (2023). La distinción entre las reflectancias planetarias (R y RC), radica en la manera en que son normalizadas. Para la normalización de RC se utiliza la expresión de la masa de aire de Young (1994), lo que produce mejores resultados al inicio y final del día que la normalización simple por coseno del ángulo cenital, utilizada para R. La diferencia entre N1 y N2 está en el modelo de brillo de fondo utilizado. El valor de N1 fue calculado con el modelo de fondo original de Alonso-Suárez et al. (2011) utilizado desde la generación vieja de satélites GOES. En cambio, el N2 se calculó con un modelo de fondo actualizado y ajustado al satélite GOES16. Algunas pruebas empíricas sugieren que este último presenta mejoras en comparación con el primero.

Las variables satelitales FR, R, RC, N1 y N2 se consideran a distintas escalas espaciales con nomenclatura del 01 al 20: FR01-FR20, R01-R20, RC01-RC20, N1_01-N1_20, N2_01-N2_20, siendo promedios espaciales en celdas cuadradas de dimensiones crecientes. El espaciado entre tamaños no sigue una relación lineal; en los tamaños más pequeños, el espaciamiento es más detallado, y viceversa para los tamaños mayores.

Modelos de aprendizaje supervisado. En continuación al estudio previo, se amplió el conjunto de modelos de aprendizaje automático supervisado. Además de la regresión lineal, RN (100 neuronas ocultas y función de activación ReLu) y el RF (30 estimadores) que fueron abordados en el trabajo anterior, se incorpora ahora el modelo de GB. Este nuevo modelo busca enriquecer el análisis al ofrecer una perspectiva adicional en la estimación de la radiación solar.

RESULTADOS

Ajuste local de los modelos CIM-ESRA y CAMS

Se descargaron los estimativos de GHI de los modelos satelitales CIM-ESRA y Heliosat-4 para el periodo 2019-2021 para la estación Luján de cada sitio web http://les.edu.uy/online/stack-loc/ (CIM-ESRA, portal LES) y https://www.soda-pro.com/web-services (Heliosat-4, portal CAMS). Los estimativos de CIM-ESRA están disponibles en una escala 10-minutal para diferentes estaciones latinoamericanas una de las cuales corresponde a Luján, Argentina. Los estimativos de CAMS no están disponibles en la resolución temporal 10-minutal trabajada en este artículo, por lo que se descargaron datos minutales y luego se integraron a la escala 10-minutal. Estos modelos se utilizaron para comparar sus indicadores de desempeño respecto a las medidas en tierra con los de los modelos estadísticos desarrollados en este trabajo.

La adaptación local de los modelos CIM-ESRA y CAMS fue ejecutada mediante cuatro enfoques distintos (Salazar et al., 2021). Las tres primeras estrategias involucraron diversas combinaciones de ajustes mediante regresión lineal simple, mientras que la cuarta opción se basó en un enfoque de mapeo

cuantílico. La disparidad entre las metodologías lineales y el mapeo cuantílico radica en su enfoque para abordar el sesgo promedio de las estimaciones. Las estrategias lineales buscan mitigar el sesgo promedio ajustando una regresión de primer orden que relaciona las estimaciones con las mediciones reales. En contraste, el enfoque del mapeo cuantílico modifica las estimaciones satelitales para lograr una mayor aproximación de la función de probabilidad acumulada (CDF) a la de los datos medidos. La elección de la adaptación específica se basó en la optimización de las métricas de rendimiento resultantes de cada ajuste para los respectivos modelos. Cabe mencionar que este artículo no tiene como objetivo abordar en detalle este punto.

El desempeño de estos modelos (respecto de las mediciones en tierra) se analizó con las siguientes métricas: $MBE, RMSE, MAE$ y R^2 (MBE es el desvío promedio (o sesgo), $RMSE$ es el error cuadrático medio, MAE es el error absoluto medio y R^2 el coeficiente de determinación). Para los primeros tres se informan también sus correspondientes valores relativos como porcentaje de la media de las medidas terrestres, los cuales nombramos respectivamente: $MBEn, RMSEn$ y $MAEn$. El valor de normalización en este trabajo es de 420,3 W/m^2. Los resultados de la evaluación de estos modelos con y sin su ajuste local se pueden ver en la Tabla 3.

Tabla 3: Resultados de las métricas de desempeño de los modelos CAMS y CIM-ESRA con y sin adaptación al sitio respecto de las mediciones en tierra. Las medidas $MBE, RMSE$ y MAE están medidas en W/m^2

	MBEn	RMSE	RMSEn	MAE	MAEn	R^2
CAMS sin adaptar	-0,99	93,38	22,22	54,94	13,07	0,894
CAMS adaptado	**0**	**91,77**	**21,83**	**55,39**	**13,68**	**0,886**
CIM-ESRA sin adaptar	1,74	76,12	17,32	50,06	11,39	0,926
CIM-ESRA adaptado	**0**	**75,29**	**17,13**	**49,31**	**11,22**	**0,931**

Se observa que los estimados del modelo CIM-ESRA demuestran una adaptación más precisa a la región en comparación con los resultados obtenidos mediante el modelo Heliosat-4. La superioridad del desempeño de CIM-ESRA se origina por dos razones: en primer lugar, hace uso de información del satélite GOES-16 en lugar de MSG, lo cual ofrece ángulos de visión más favorables para la región de la Pampa Húmeda. En segundo lugar, CIM-ESRA se caracteriza por ser un modelo semi-empírico, cuyos parámetros ajustables fueron determinados específicamente para la región en base a datos de 10 ubicaciones durante el periodo 2010-2017. En consecuencia, la referencia de rendimiento para este trabajo la establece en el modelo CIM-ESRA, el cual, además, emplea información del mismo satélite que los datos de entrada utilizados en los algoritmos de aprendizaje automático presentados en este estudio. Por otro lado, se evidencia que entre las métricas examinadas, aquella que muestra una mejora es el sesgo, el cual tiende a aproximarse a 0. Mientras tanto, las demás métricas tienden a mejorar levemente, salvo en el caso del MAE del modelo CAMS donde se observa una pequeña desmejoría. En cualquier caso, la ganancia observada de los métodos de post-proceso para adaptación al sitio es limitada, estando en general por debajo del 1%.

Implementación del modelo CIM-McClear

Se implementó el modelo CIM-McClear con el propósito de contar con una referencia adicional, más exigente, para comparar el rendimiento frente a de los modelos de aprendizaje automático. Además, su forma de ajuste y testeo es la misma que la utilizada para los algoritmos de aprendizaje automático. Los modelos pertenecientes a la familia CIM (Cloud Index Model) se caracterizan por una estructura que combina un modelo de cielo despejado con un factor de atenuación que considera el efecto de las nubes. Este factor de atenuación, denotado como F, se rige por una función lineal que se vincula al índice de nubosidad derivado de datos satelitales como se ve en la ecuación (2) y (3).

$$GHI = GHI_{CS} * (a + b(1 - N)) \quad (2)$$

$$F(N) = a + b(1 - N) \quad (3)$$

donde a y b son parámetros que se ajustan localmente, GHI_{CS} es la irradiancia en condiciones de cielo despejado (en este artículo se utilizó el modelo McClear) y N es el índice de nubosidad. Se utilizó en este caso el índice N1, según la implementación usual. Para realizar la implementación, se tomaron los años 2019, 2020 y 2021, asignando dos de ellos para el entrenamiento y reservando el restante para la validación. El índice de nubosidad N corresponde a una resolución espacial de 100 x 100 (variable N1_08) la cual resultó ser la óptima. La figura 1 muestra la distribución y densidad de los puntos, comparando los valores reales con sus respectivas predicciones. Los promedios de las métricas obtenidas durante estos tres años se presentan en la tabla 4.

Tabla 4: Resultados de los promedios de las métricas de desempeño del modelo implementado CIM-McClear, las medidas MBE, RMSE y MAE están medidas en W/m2

	MBEn	**RMSE**	**RMSEn**	**MAE**	**MAEn**	**R^2**
CIM-McClear implementado	0	70,74	16,15	42,37	13,07	0,943

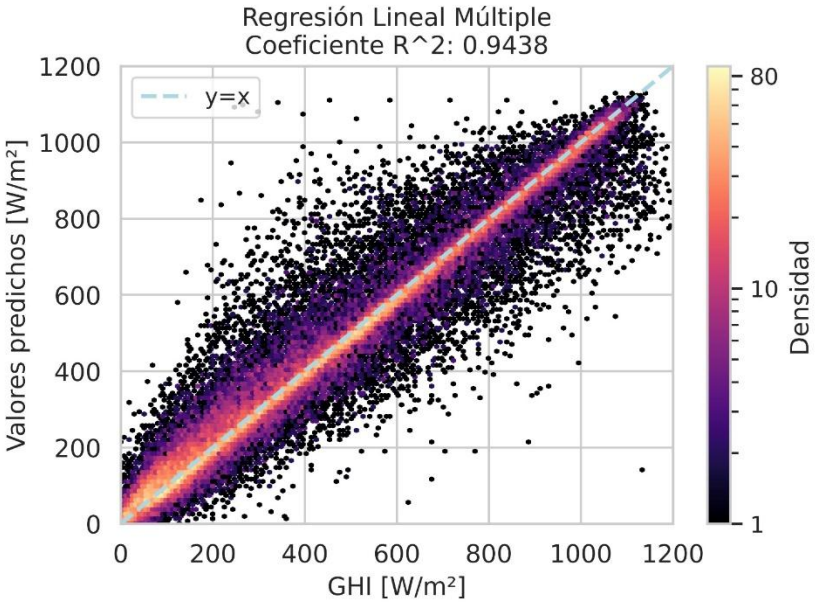

Figura 1: Distribución y densidad de los puntos para los valores reales vs. valores predicho.

Implementación de los modelos de aprendizaje automático

Se utilizaron como entrada las variables mencionadas en la sección *Modelos estadísticos con aprendizaje automático*. Para los modelos implementados se probaron diferentes combinaciones de las variables de entrada, encontrándose que para los algoritmos de aprendizaje automático fue suficiente contar con la información de las variables provenientes de imágenes satelitales. Se presentan las métricas para cada año de validación, siendo el ajuste con los otros dos años que completan el periodo 2019-2021. En la última columna se presenta el promedio de desempeño entre los 3 años de validación, lo que se usa como valor de comparación con la referencia del CIM-ESRA de la Tabla 3.

El primer objetivo fue saber si las correcciones en el índice de nubosidad N1 que da lugar al índice N2 y la corrección a la reflectancia planetaria R que da lugar a la reflectancia RC llevan a mejoras en los resultados. Con las variables disponibles se armaron 4 distintos conjuntos de datos para encontrar la combinación de variables más adecuada. Cabe aclarar que cada conjunto contiene todo el set de variables de cada tipo en sus distintas resoluciones.

Conjunto 1: R - FR - N1

Conjunto 2: R - FR - N2

Conjunto 3: RC- FR - N1

Conjunto 4: RC - FR - N2

Se aplicaron los modelos de aprendizaje automático RF, GB y RN. El desempeño fue muy similar y no permitió concluir qué conjunto es más adecuado. El mejor desempeño se obtuvo con RN, seguido de RF y por último GB, como se muestra en la Figura 2. Con RN y RF se logró en todos los casos una pequeña mejoría respecto a los resultados (Iturbide, et al. 2023) donde las variables utilizadas fueron FR y R.

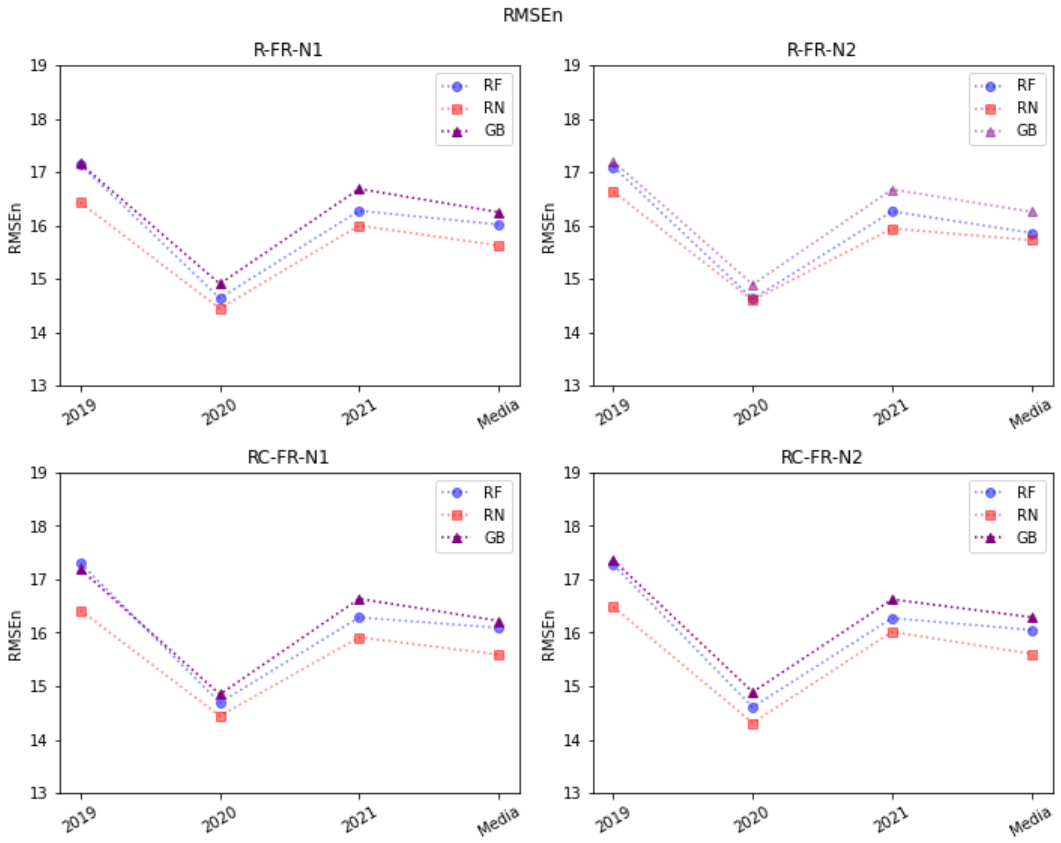

Figura 2: Error cuadrático medio porcentual para los distintos modelos empleados

Reducción de dimensionalidad del conjunto de datos

El conjunto de variables N1, N2, R y RC está altamente correlacionado. Se realizó un análisis de componentes principales para disminuir la dimensión de los conjuntos, reteniendo hasta la componente principal 5 inclusive. En todos los conjuntos el porcentaje de varianza acumulada con estas 5 componentes es mayor al 99.5% y se recuperan los valores RMSEn obtenidos al trabajar con todas las variables (diferencias menores al 0,05%). Tomar más componentes principales no mejora los desempeños (ver tabla 4).

Se buscaron qué variables eran las más relevantes para el modelo. Para eso, se analizó primero qué variable presenta mayor correlación con la variable a predecir. En todos los casos (FR, R, RC, N1 y N2) estas variables se encuentran alrededor de la resolución media (variables 9 o 10). Se buscaron 2 variables

más en los extremos de las resoluciones. La búsqueda no fue exhaustiva y se encontraron mejores resultados con RN que con RF. Cuando se reduce la dimensión del conjunto de esta manera, RF aumenta 0,4% aproximadamente respecto al desempeño del algoritmo sobre el conjunto de datos que contiene a todas las variables, mientras que la RN da valores más cercanos y aumenta 0,2% como máximo. Para mejorar el desempeño se agregó la variable geométrica cosz. Con la referencia geométrica los dos algoritmos se aproximan a los valores de desempeño cuando se utilizan todas las variables.

Los resultados de los desempeños promediados en los 3 años se muestran en la Figura 3 para los algoritmos RF y RN en los conjuntos de datos formados con: todas las variables (A), las 5 primeras componentes principales (B), una selección de variables (C) y la selección anterior más cosz (D). En particular se muestran los resultados del conjunto 4: RC - FR - N2. La selección C corresponde a FR2, FR10, FR14, N2_1, N2_9, N2_14, RC1, RC10 y RC14.

Tabla 4: Resultados de los modelos de ML utilizando las variables FR-RC-N2.

	MBEn	**RMSE**	**RMSEn**	**MAE**	**MAEn**	**R^2**
RF A	-0,08	70,25	16,05	42,86	9,80	0,939
RN A	0,03	68,32	15,60	41,88	9,57	0,943
RF B	-0,07	70,29	16,06	42,92	9,81	0,939
RN B	0,05	68,14	15,56	41,44	9,47	0,944
RF C	-0,04	71,92	16,42	43,84	10,02	0,937
RN C	-0.34	69,01	15,75	41,29	9,43	0,943
RF D	0,05	70,11	16,01	41,95	9,59	0,941
RN D	0,04	68,65	15,67	41,78	9,55	0,943

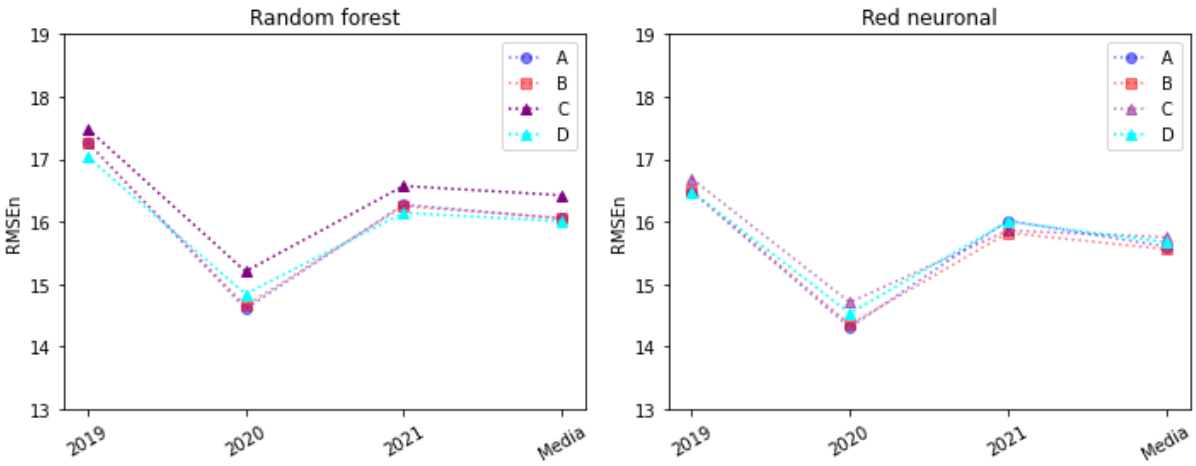

Figura 3: Error cuadrático medio porcentual para los algoritmos RF y RN de las 4 selecciones de datos utilizadas para las variables RC - FR - N2

Visualización comparativa de estimaciones

Las Figuras 4 y 5 exhiben las proyecciones generadas por la RN (representadas por la línea roja) en contraste con los valores registrados en tierra (indicados por la línea punteada), junto con la estimación obtenida mediante la implementación del modelo CIM-McClear, que se distinguió como la más precisa

en términos de comparación, para días con nubosidad y claros, respectivamente. Existe notoria concordancia entre las aproximaciones, aunque se aprecian sutiles diferencias en las mismas. Tanto la RN como el modelo CIM-McClear logran de manera efectiva capturar las condiciones de nubosidad y la GHI en términos generales.

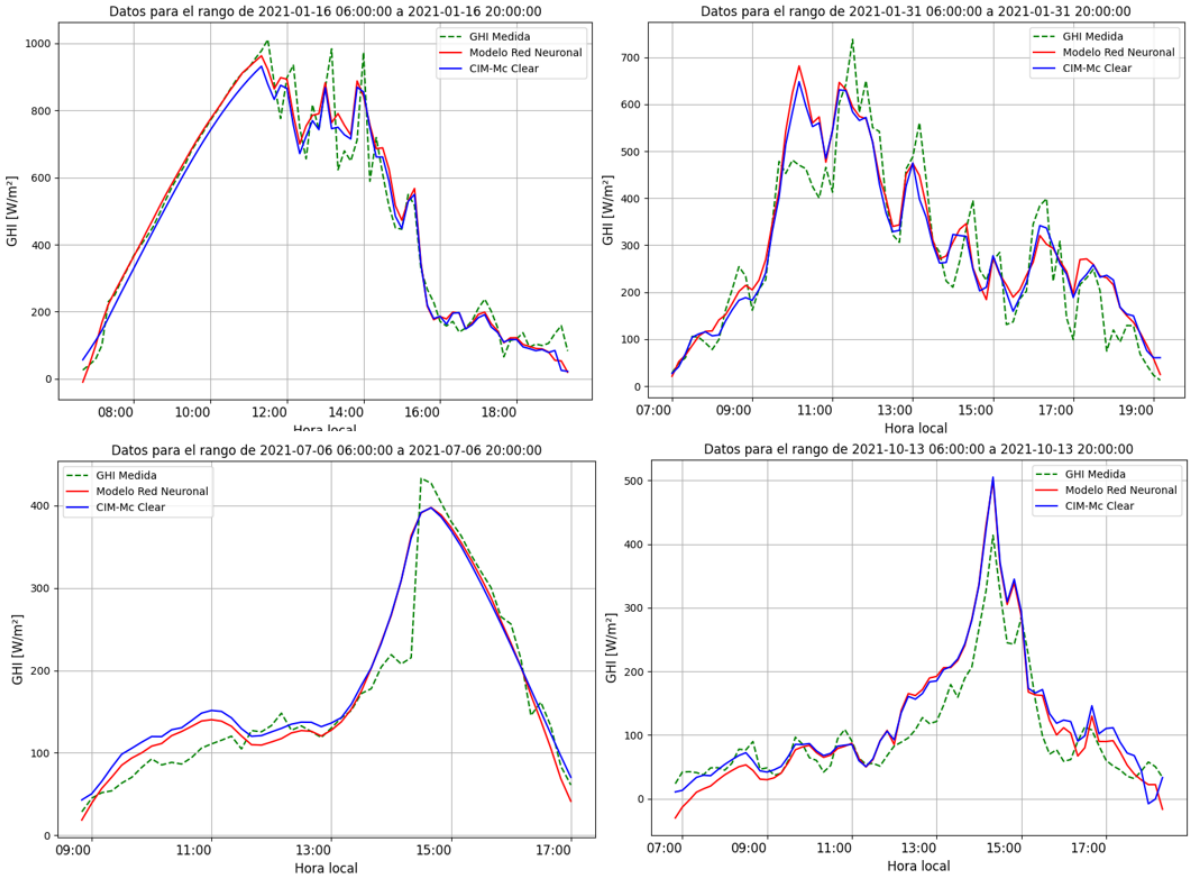

Figura 4: Gráficos comparativos entre las medidas de tierra, el modelo CIM-McClear implementado y la red neuronal para cuatro días de 2021 en condiciones de nubosidad.

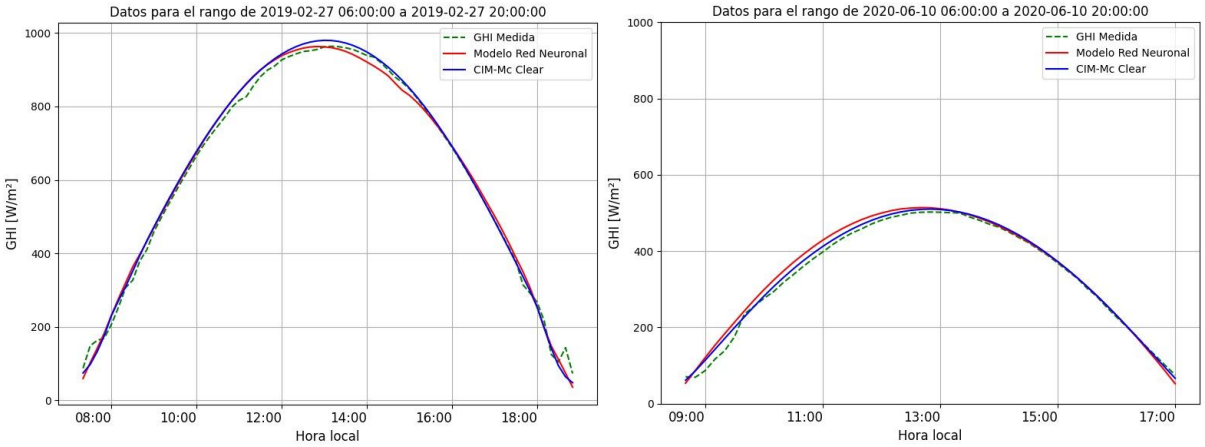

Figura 5: Gráficos comparativos entre las medidas de tierra, el modelo CIM-McClear implementado y la red neuronal para dos días despejados de 2021.

CONCLUSIONES

Entre los diversos algoritmos de Aprendizaje Automático utilizados en este estudio, se observó un mejor rendimiento por parte de la RN, seguido del método de RF, mientras que el método de GB quedó

rezagado. La combinación de variables FR, RC y N2 mostró los resultados más favorables, obteniendo un error cuadrático medio promedio porcentual del 15,56% después de la reducción de la dimensionalidad mediante el análisis de componentes principales. Al comparar con el modelo de referencia CIM-ESRA, los modelos empíricos de ML demostraron un rendimiento superior. Las métricas comparativas entre CIM-ESRA y la RN favorecieron a esta última: MAEn de 9,7% vs. 11,3%, RMSEn de 15,6% vs. 17,1%.

Una comparación interesante surge con el modelo CIM-McClear, que logró un RMSEn promedio de 16,15% y un sesgo nulo. Cabe destacar que este modelo requiere solo la entrada del modelo de cielo claro y una variable de nubosidad parametrizada mediante una función lineal. Su simplicidad ofrece un rendimiento menor pero similar al aprendizaje automático, que además utiliza información espacial multiescala.

El análisis de componentes principales no redujo de manera significativa los errores. En resumen, el modelo de ML propuesto mejora las estimaciones para la región, superando a modelos ajustados al sitio y al CIM-McClear ajustado localmente, con métricas de desempeño levemente mejores.

En futuros trabajos se debe analizar el comportamiento del modelo empírico en otras áreas de la Pampa Húmeda, extrapolando espacialmente mediante pruebas en una ubicación distinta. También sería recomendable considerar otras variables satelitales relevantes, como los canales infrarrojos del satélite, que brindan información adicional sobre el sistema Tierra-Atmósfera y podrían mejorar la precisión de las estimaciones de radiación solar. Evaluar la inclusión de este canal en el modelo y su impacto en el rendimiento sería valioso.

AGRADECIMIENTOS

R. Alonso-Suárez agradece a la Comisión Sectorial de Investigación Científica (CSIC), Udelar, por el apoyo financiero al Laboratorio de Energía Solar a través de su programa de Grupos de I+D.

Los investigadores del GERSolar agradecen a la Secretaría de Ciencia y Tecnología (SCyT) de la UNLu por el apoyo financiero a través de los Proyectos de Investigación para Investigadores en Formación 2022 (RESREC-LUJ: 266-22) y al Ing. Lucas Burgos por la recopilación y el control de calidad realizado a los datos de la estación Luján utilizados en este trabajo.

REFERENCIAS

Abal, G., Aicardi, D., Alonso-Suárez, R., y Laguarda, A. (2017). Performance of empirical models for diffuse fraction in uruguay. Solar Energy, 141:166–181.

Alonso-Suárez, R., Abal, G., Siri, R., y Musé, P. (2012). Brightness-dependent Tarpley model for global solar radiation estimation using GOES satellite images: application to Uruguay. Solar Energy, 86 , 3205–3215. doi:10.1016/j.solener.2012.08.012.

Gonzalez, J., Teixeira-Branco, V., y Alonso-Suárez, R. (2019). Evaluation of the Heliosat-4 and FLASH-Flux models for solar global daily irradiation estimate in Uruguay. En ISES Conf. Proceedings, Solar World Congress.

Iturbide, P., Alonso-Suarez, R., Ronchetti, F. (2023). An Analysis of Satellite-Based Machine Learning Models to Estimate Global Solar Irradiance at a Horizontal Plane. In: Naiouf, M., Rucci, E., Chichizola, F., De Giusti, L. (eds) Cloud Computing, Big Data & Emerging Topics. JCC-BD&ET 2023. Communications in Computer and Information Science, vol 1828. Springer, Cham. https://doi.org/10.1007/978-3-031-40942-4_9

Jiménez, V. A., Will, A., & Rodríguez, S. (2017). Estimación de radiación solar horaria utilizando modelos empíricos y redes neuronales artificiales. Ciencia y tecnología, (17), 29-45.

Laguarda, A., Iturbide, P., Orsi, X., Denegri, M. J., Luza, S., Burgos, B. L., Stern, V., y Alonso-Suárez, R. (2021). Validación de modelos satelitales Heliosat-4 y CIM-ESRA para la estimación de irradiancia solar en la Pampa Húmeda. Energías Renovables y Medio Ambiente, 48, 1-9.

Laguarda, A., Giacosa, G., Alonso-Suárez, R., y Abal, G. (2020). Performance of the site-adapted CAMS database and locally adjusted cloud index models for estimating global solar horizontal irradiation over the Pampa Húmeda region. Solar Energy, 199:295–307.

Laguarda, A., Alonso-Suárez, R., y Abal, G. (2018). Modelo semi-empírico de irradiación solar global a partir de imágenes satelitales GOES. Anales del VII Congresso Brasileiro de Energia Solar.

Lefèvre, M., Oumbe, A., Blanc, P., Espinar, B., Qu, Z., Wald, L., Homscheidt, M. S., y Arola, A. (2013). McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmospheric Measurement Techniques, European Geosciences Union, 6 , 2403–2418. doi:10.5194/amt-6-2403-2013.

Long, C. N., & Shi, Y. (2008). An automated quality assessment and control algorithm for surface radiation measurements. The Open Atmospheric Science Journal, 2(1).

McArthur, L. (2005). Baseline Surface Radiation Network (BSRN) Operations Manual. Td-no. 1274, wrcp/wmo, World Meteorological Organization (WMO, www.wmo.org).

Olivera, L., Atia, J., Amet, L., Osio, J., Morales, M., & Cappelletti, M. (2020). Uso de redes neuronales artificiales para la estimación de la radiación solar horaria bajo diferentes condiciones de cielo. Avances en Energías Renovables y Medio Ambiente-AVERMA, 24, 232-243.

Perez, R., Ineichen, P., Seals, R., & Zelenka, A. (1990). Making full use of the clearness index for parameterizing hourly insolation conditions. Solar Energy, 45(2), 111-114.

Perez, R., Cebecauer, T., & Šúri, M. (2013). Semi-empirical satellite models. Solar energy forecasting and resource assessment, 21-48.

Qu, Z., Oumbe, A., Blanc, P., Espinar, B., Gesell, G., Gschwind, B., Klüser, L., Lefèvre, M., Saboret, L., Schroedter-Homscheidt, M., y Wald, L. (2017). Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method. Meteorologische Zeitschrift, 26(1):33–57.

Raichijk, C. (2008). Estimación de la irradiación solar global en Argentina mediante el uso de redes neuronales. Energías Renovables y Medio Ambiente (ISSN 0328-932X). Vol. 22, pp. 1 - 6.

Salazar, G. A., Alonso-Suárez, R., Cirigliano, A. L., y Ledesma, R. D. (2021). Evaluación del proceso de adaptación al sitio aplicado a la irradiancia solar global medida en la ciudad de Salta, Argentina. Avances en Energías Renovables y Medio Ambiente-AVERMA, 25, 353-362.

Sarazola, I., Laguarda, A., Ceballos, J. C., y Alonso-Suárez, R. (2023). Benchmarking of modeled solar irradiation data in Uruguay at a daily time scale. IEEE Latin American Transactions.

Sayago, S., Bocco, M., Ovando, G., & Willington, E. A. (2011). Radiación solar horaria: modelos de estimación a partir de variables meteorológicas básicas. Avances en Energías Renovables y Medio Ambiente, 15.

Verbois, H., Saint-Drenan, Y.-M., Becquet, V., Gschwind, B., Blanc, P. (2023). Retrieval of surface solar irradiance from satellite using machine learning: pitfalls and perspectives, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2023-243.

Yang, D. (2020). Choice of clear-sky model in solar forecasting. Journal of Renewable and Sustainable Energy 12, 026101, https://doi.org/10.1063/5.0003495.

Young, A.T. (1994). Air mass and refraction. Applied optics, 33 6, 1108-10 .

MACHINE LEARNING MODELS FOR ESTIMATING SOLAR HORIZONTAL RADIATION IN THE PAMPA HÚMEDA WITH MULTISCALE SATELLITE INFORMATION

ABSTRACT: The lack of precision in solar radiation data impacts the solar energy projects risk. Ground measurement networks provide limited information due to their sparse spatial distribution. This leads to estimation models based on satellite imagery, solving the spatial issue if carefully adjusted to quality ground measurements. In this article, we develop and validate an empirical Machine Learning (ML) model for satellite-based solar radiation estimation, demonstrating its usefulness and accuracy in the studied region. The models are fed with variables from GOES-16 satellite imagery, McClear model estimates, and geometric data. Our results suggest that for certain proposed models, satellite information is sufficient for accurately estimating solar radiation, by obtaining the temporal reference from implicit relationships between the considered satellite variables. Given the size of the data set, we propose a principal component analysis to reduce dimensionality. In order to compare the proposed model, we adapt Heliosat-4 and CIM-ESRA estimates to the site and implement the CIM-McClear model. The

results indicate that the proposed model outperforms others, although slightly, showing how difficult it is to further improve solar radiation satellite-based estimation.

Keywords: Solar radiation, Machine Learning, Satellite images, GOES16, GHI.