



UNIVERSIDAD NACIONAL DE TUCUMAN

FACULTAD DE CIENCIAS EXACTAS Y TECNOLOGIA

Aplicaciones de métodos de aprendizaje automático al estudio de la ionosfera

Tesis de Licenciatura en Física

Autor:

Isaías Emanuel Nallar

Director:

Bruno S. Zossi

Co-Director:

Ana G. Elias

Tucumán-Argentina

Junio de 2026

Contenido

Contexto personal y agradecimientos.....	3
Resumen.....	4
Introducción	5
Capítulo 1. La ionosfera.....	6
1.1 Introducción histórica	6
1.2 Descripción general de la atmósfera	7
1.3 Densidad electrónica en la ionosfera y procesos físicos que la afectan	10
1.4 Propagación de ondas de radio en un medio ionizado	11
1.4.1 Teoría magneto-iónica y la ecuación de Appleton.....	11
1.4.2 Reflexión ionosférica de ondas de alta frecuencia.....	12
1.5 Deducción de la frecuencia de plasma.....	13
1.6 Técnicas de observación del entorno espacial de la Tierra.....	14
1.6.1 Importancia de las observaciones y clasificación de las técnicas de sensado	15
1.6.2 Sondeo ionosférico por ondas de radio	15
1.6.3 Ionosonda.....	16
Capítulo 2. Variabilidad temporal del parámetro ionosférico foF2 y sus principales forzantes.....	20
2.1 Variabilidad espacial de la capa F2.....	20
2.2 Variación diurna	21
2.3 Variación estacional	21
2.4 Variación con la actividad solar	22
2.5 Proxies de la actividad solar	23
2.5.1 Número de manchas solares, SN.....	24
2.5.2 Índice MgII	25
2.5.3 Flujo en onda de radio F10.7	26
2.5.4 Flujo en onda de radio F30.....	27
2.6 Variación con la actividad geomagnética	27
2.7 Índices de la actividad geomagnética.....	28
2.8 Índice global ionosférico IG	30
Capítulo 3. Métodos de interpolación basados en ML	32
3.1 ¿Qué es el "Machine Learning" (ML)?.....	32
3.2 Regresión lineal	33
3.3 Regresión por Bosques Aleatorios (Random Forest Regression)	34
3.4 Regresión por Gradient Boosting (Gradient Boosting Regression)	35
3.5 Redes Neuronales (Neural Networks).....	37
Capítulo 4. Datos y Metodología.....	40

4.1 Fuentes de datos de foF2	40
4.2 Estación seleccionada: Juliusruh	43
4.3 Generación de huecos.....	47
4.4 Evaluación del desempeño de los modelos	49
4.4.1 Coeficiente de determinación (R^2)	50
4.4.2 Error Absoluto Medio (MAE).....	50
4.4.3 Error Absoluto Medio Porcentual (MAPE)	50
Capítulo 5. Resultados.....	52
5.1 Interpolación de foF2 utilizando regresión lineal.....	52
5.2 Interpolación de foF2 utilizando Bosques Aleatorios (Random Forest Regression, RFR).....	60
5.3 Interpolación de foF2 utilizando Gradient Boosting Regression (GBR)	63
5.4 Interpolación de foF2 utilizando Redes Neuronales	65
5.5 Análisis comparativo del desempeño de cada método de interpolación	66
Capítulo 6. Discusión y conclusiones.....	71
6.1 Método de interpolación más conveniente en el caso del parámetro ionosférico foF2: simpleza vs. precisión.....	71
6.2 Aportes de los resultados de esta Tesis a la problemática actual de las series de datos con huecos y a la necesidad de series completas	73
6.3 Comparación con otros trabajos que utilizan técnicas de ML para predecir valores de parámetros ionosféricos	73
6.4 Líneas futuras de investigación	74
Referencias	76

Contexto personal y agradecimientos

Contexto personal

Comencé la carrera de Licenciatura en Física a los 27 años, siendo abogado recibido en la Universidad Nacional de Tucumán. Tomar esa decisión no fue fácil: implicaba restar tiempo a mi trabajo, renunciar a mis ingresos y volver a volcarme por completo al estudio. Fue un camino lleno de desafíos.

Recién abandoné mi trabajo cuando estaba en cuarto año de la carrera. Necesité mucho orden, disciplina y convicción para sostener este rumbo.

Mi entorno nunca entendió qué es lo que yo buscaba o por qué lo hacía. Ya había invertido largos años en obtener un título y en vez de abocarme a trabajar con el, inicié otra carrera que nada tenía que ver con la anterior. Siempre me sentí incomprendido y solo en este nuevo objetivo, aunque ello no resultó un obstáculo para mí.

Tenía claro que no quería transitar la vida sin comprender las leyes fundamentales que rigen el universo. Para mí, esta carecía de sentido sin ese conocimiento.

Cada vez que leo los textos que los grandes científicos de la historia nos legaron siento que de alguna forma estoy dialogando con ellos y ello me llena de placer. Cada día reflexiono sobre los misterios del universo y la existencia de la mano de las mentes mas brillantes, que compartieron su conocimiento con todos nosotros. No tenía dudas de que ese es el lugar donde yo quería estar. Me emociona pensar que ellos sintieron la misma curiosidad incontenible, a veces casi con una obsesión enfermiza, sumado a su genio incomparable, que nos permite apreciar los secretos más sutiles de la existencia, escritos en el código matemático.

Debo ser honesto, a veces me siento más identificado con la filosofía que con la física. De alguna manera esta es, en mi opinión, la verdadera filosofía.

Quizá la verdad sea, que soy una persona del humanismo que decidió estudiar física. No puedo negar mis orígenes en humanidades y que estos hayan influido y lo hagan actualmente en la perspectiva a la hora de mirar el mundo.

Agradecimientos

Agradezco al Director de Tesis Bruno Zossi. A la Co-Directora Ana Georgina Elias, que más allá de lo estrictamente académico mostró una gran empatía a nivel humano, habilidades de conducción y liderazgo (soft skills). A mi Tío, Ing. Pablo Viviani, que además del afecto me ayudó económicamente para poder abocarme al estudio.

A los profesores que, cada uno a su manera, me transmitieron pasión y conocimiento. A los compañeros que conocí y cursaron conmigo. Fue una experiencia hermosa hacer esta carrera.

Al Departamento de Física de la Facultad de Ciencias Exactas y Tecnología de la Universidad Nacional de Tucumán, a la Facultad y a la Universidad.

Resumen

El objetivo principal de este trabajo de tesis es evaluar y contrastar el desempeño de distintos métodos de aprendizaje automático (Machine Learning, ML) aplicados a la interpolación de datos faltantes en series temporales de la frecuencia crítica de la capa F2 (f_oF2). El análisis abarca un espectro progresivo de complejidad algorítmica: comenzando con un enfoque clásico de regresión lineal, continuando con modelos basados en ensambles como Random Forest Regressor y Gradient Boosting Regressor, hasta arribar a la implementación de redes neuronales.

Para someter a prueba la capacidad de generalización de estos modelos, se diseñaron diferentes escenarios de pérdida de información sobre los conjuntos de datos. Estos incluyeron, por un lado, ventanas temporales con ausencias continuas y prolongadas de datos y, por otro lado, remociones aleatorias con tasas de pérdida que oscilaron entre el 60% y el 97% del total de la serie.

La evaluación de los modelos en condiciones diversas permitió analizar su comportamiento, posibilitando la comparación de sus metodologías de funcionamiento, así como la identificación de sus fortalezas, debilidades y criterios de selección según el contexto operativo. Más allá de la ejecución técnica y la implementación del código en el lenguaje de programación, el núcleo de esta investigación radica en comprender de manera profunda los principios matemáticos y algorítmicos que rigen a cada predictor. Por esta razón se mantuvo la simplicidad de los modelos en su programación, evitando añadir funciones preestablecidas que pudieran operar como cajas negras o cuyo nivel estuviere fuera del alcance de la profundidad pretendida para este trabajo.

Asimismo, este análisis metodológico se ensambla con el estudio de la física ionosférica subyacente a los datos utilizados, garantizando que las soluciones e interpretaciones de los modelos de aprendizaje automático guarden una estricta coherencia con los fenómenos geofísicos que gobiernan la dinámica del plasma ionosférico.

Introducción

El Machine Learning (ML), o Aprendizaje Automático, es una disciplina de la Inteligencia Artificial (IA). Se refiere a las tecnologías y algoritmos que permiten a los sistemas identificar patrones, tomar decisiones y mejorar a través de la experiencia y los datos. Las técnicas de ML han demostrado ser muy útiles para abordar problemas de Física Espacial logrando excelentes resultados. En particular, uno de los problemas más demandantes, además de los pronósticos a tiempo futuro, es el llenado de huecos en series de datos de mediciones a partir de distintos instrumentos (in situ y satelitales) para estudiar fenómenos ionosféricos con cobertura de datos completa en el tiempo y en el espacio.

En este trabajo se explora el uso de algoritmos de ML para completar bases de datos de la frecuencia crítica de la capa F2 de la ionosfera, foF_2 , que es una medida de la densidad electrónica máxima (NmF_2) de esta región de la atmósfera. Este parámetro ionosférico es uno de los más importantes para aplicaciones de meteorología espacial, posicionamiento preciso y telecomunicaciones, por lo que resulta esencial tanto para estudios científicos como para aplicaciones tecnológicas.

Entrenar un modelo basado en ML no es una tarea trivial ya que, si bien existen códigos desarrollados por la comunidad que facilitan la tarea, crear un modelo requiere un conocimiento profundo de los datos utilizados.

En este trabajo se comienza con una descripción general de la ionosfera, junto con un análisis de la variabilidad de foF_2 y sus forzantes. Entre estos últimos se analiza en particular el forzante solar. Se presenta luego el enfoque computacional adoptado, detallando las técnicas de programación utilizadas y, en particular, los métodos aplicados al interpolado de datos, entre los que se incluyen métodos basados en ML: regresión lineal simple y múltiple, Gradient Boosting, Random Forest y Redes Neuronales. Se presentan luego los datos empleados, la metodología de análisis y los resultados obtenidos para cada uno de los métodos aplicados. Finalmente, se ofrece una valoración comparativa del desempeño de cada técnica, se discuten los resultados obtenidos junto con la importancia de los mismos, destacando los aportes, limitaciones y posibles líneas de investigación futura.

Capítulo 1. La ionosfera

En este capítulo se describen las características principales de la atmósfera y conocimientos básicos sobre la ionosfera dado que la investigación desarrollada en esta Tesis se focaliza en un parámetro de particular relevancia en esta región atmosférica: la frecuencia crítica de la capa F2 (foF2). Precisamente, el tema de investigación se centra en métodos de interpolación de datos faltantes de foF2, los cuales requieren de cierto conocimiento de las características de variabilidad de este parámetro. Se incluye también las técnicas de medición de foF2.

1.1 Introducción histórica

El primer indicio de la existencia de una capa conductora en la atmósfera terrestre estuvo vinculado al estudio del geomagnetismo. Desde el siglo XVIII se conocían pequeñas variaciones diarias del campo magnético, y ya en 1839 Carl Friedrich Gauss especuló que podían tener su origen en corrientes eléctricas atmosféricas. Posteriormente, Lord Kelvin (1860) sugirió la posibilidad de que existiera una capa conductora asociada a fenómenos de electricidad atmosférica. Si en cierto nivel de la atmósfera el aire resultaba suficientemente conductor, los vientos al cruzar el campo magnético terrestre generarían campos y corrientes eléctricas (Rishbeth & Garriott, 1969). Estas variaciones se explicarían entonces con “teoría del dínamo”.

La teoría de una “capa conductora” adquirió renovada importancia en 1901, cuando Guglielmo Marconi logró transmitir señales de radio a través del Atlántico. Este resultado implicaba que las ondas de radio eran desviadas en un grado mayor al esperado por simple difracción. Y, en 1902, Arthur Edwin Kennelly y Oliver Heaviside propusieron que la presencia de cargas libres en la atmósfera superior podía reflejar dichas ondas (Rishbeth & Garriott, 1969). El estudio de la ionósfera surgió, entonces, de la necesidad de comprender los efectos de la atmósfera superior ionizada sobre la propagación de las ondas de radio. Fue a partir de las observaciones de estas ondas que se confirmó la existencia de una región reflectora compuesta por electrones e iones, capaz de explicar fenómenos que no podían atribuirse a la atmósfera neutra. Desde entonces, la ionósfera se convirtió en objeto de intensa investigación, en particular sobre la formación y variabilidad de sus diferentes capas, las cuales muestran dependencia con la hora local, la posición geográfica, la estación del año y otros factores geofísicos.

En las primeras décadas del siglo XX, el desarrollo de la ionosonda, un instrumento de detección remota que permite obtener perfiles de densidad electrónica hasta el máximo local de ionización, marcó un hito en la investigación ionosférica. A través de este tipo de mediciones, se puso en evidencia una gran diversidad de comportamientos, desde perfiles reproducibles y tranquilos hasta otros completamente caóticos. Además, se observaron variaciones tanto periódicas como aperiódicas con escalas temporales que van desde segundos hasta décadas, como es el caso del ciclo solar de 11 años.

Con el tiempo, el enfoque de la investigación migró desde un interés centrado en los mecanismos de producción y pérdida de plasma hacia una perspectiva más dinámica y física, impulsada por el desarrollo de la Física de Plasmas como marco teórico. El auge del programa espacial, junto con el desarrollo de instrumentos para globos, cohetes y satélites, posibilitó mediciones in situ que complementaron a las observaciones desde Tierra, acelerando notablemente el progreso de la disciplina (Hargreaves, 1992).

En paralelo, técnicas avanzadas como el radar de dispersión incoherente permitieron acceder a mediciones detalladas de parámetros ionosféricos en altitudes antes inaccesibles, reforzando aún más la necesidad de un conocimiento profundo en Física de Plasmas. Todo este desarrollo condujo a una transformación del enfoque original: hoy, la física de la ionósfera se comprende mejor como una rama de la Física de Plasmas del espacio, en la cual confluyen también elementos de la dinámica atmosférica, la magnetosfera y la fotoquímica (Hargreaves, 1992).

1.2 Descripción general de la atmósfera

La atmósfera terrestre presenta una estructura mucho más compleja de lo que usualmente se supone. En sus niveles más bajos, constituye un gas relativamente denso compuesto principalmente por nitrógeno y oxígeno moleculares, acompañados por trazas de dióxido de carbono, vapor de agua y otros gases en menores concentraciones. Sin embargo, tanto la presión como la densidad disminuyen exponencialmente con la altitud. A modo de referencia, a 50 km de altura el 99.9% de la masa total de la atmósfera se encuentra por debajo, y a 100 km la concentración de masa cae a menos de una parte por millón.

A estas altitudes elevadas, la radiación en el rango del extremo ultravioleta (EUV) (10-120 nm) y de rayos X provenientes del Sol tienen la capacidad de penetrar y afectar significativamente a las especies atmosféricas. Estos fotones poseen suficiente energía como para disociar e ionizar los átomos y moléculas presentes, modificando así la composición química del medio y provocando ionización y calentamiento. Este proceso da lugar a una región superior denominada termosfera, caracterizada por su baja turbulencia y por la separación de gases según su peso (Rishbeth & Garriott, 1969; Hargreaves, 1992).

Por encima de los 100 km, la baja presión impide una recombinación rápida de las especies ionizadas, lo que favorece la existencia de una población persistente de iones y electrones libres. La concentración de electrones alcanza su máximo en altitudes de aproximadamente 300 km. Aunque esta concentración representa apenas el 1% de la densidad de partículas neutras, su presencia modifica de forma sustancial las propiedades físicas del medio. Esta región, conocida como ionosfera, consiste en un plasma débilmente ionizado, capaz de conducir electricidad y sostener intensas corrientes eléctricas (Rishbeth & Garriott, 1969; Hargreaves, 1992).

La ionosfera no solo tiene implicancias en la dinámica atmosférica, sino que juega un rol fundamental en la propagación de ondas de radio. Como plasma, puede albergar una variedad de modos de onda, inestabilidades y fenómenos de interacción que no están presentes en un gas neutro. Su estudio, por lo tanto, trasciende el ámbito tradicional de la meteorología y requiere una aproximación interdisciplinaria que combine Física Atmosférica, Electromagnetismo, Magnetohidrodinámica y Física del Plasma. Precisamente, se define a la ionosfera como la región de la atmósfera en la que la cantidad de iones y electrones libres es suficiente como para afectar la propagación de ondas de radio (Rishbeth & Garriott, 1969; Hargreaves, 1992; Schunk & Nagy, 2009).

Debido a la influencia gravitatoria, tanto la atmósfera como la ionosfera presentan, en primera aproximación, una estratificación horizontal. La estructura atmosférica puede describirse mediante un perfil representativo de temperatura, mientras que la ionosfera se describe en general de acuerdo al perfil de la densidad numérica del plasma, como se indica esquemáticamente en la Figura 1.

En latitudes medias, la temperatura atmosférica decrece con la altitud a razón de aproximadamente 7 K/km en la troposfera, hasta alcanzar un mínimo en la tropopausa, entre los 10 y 12 km. Luego se encuentra la estratosfera donde la temperatura aumenta con la altura debido a la absorción de radiación ultravioleta por parte del ozono, hasta un máximo a alrededor de 50 km, donde se encuentra la estratosfera. La temperatura disminuye de nuevo con la altura en esta región que se denomina mesosfera hasta un mínimo en la mesopausa, entre los 80 y 85 km. En este punto la atmósfera alcanza el punto más frío (~ 180 K en promedio). Por encima de la mesopausa, la absorción de la radiación solar ultravioleta y EUV hace que el gradiente de temperatura sea de nuevo positivo. Esta región se denomina termosfera. Luego la temperatura se vuelve casi constante. Esta es la región de mayor temperatura de la atmósfera y se extiende hasta los 600 -1000km. Luego continúa la exosfera sin un límite superior definido.

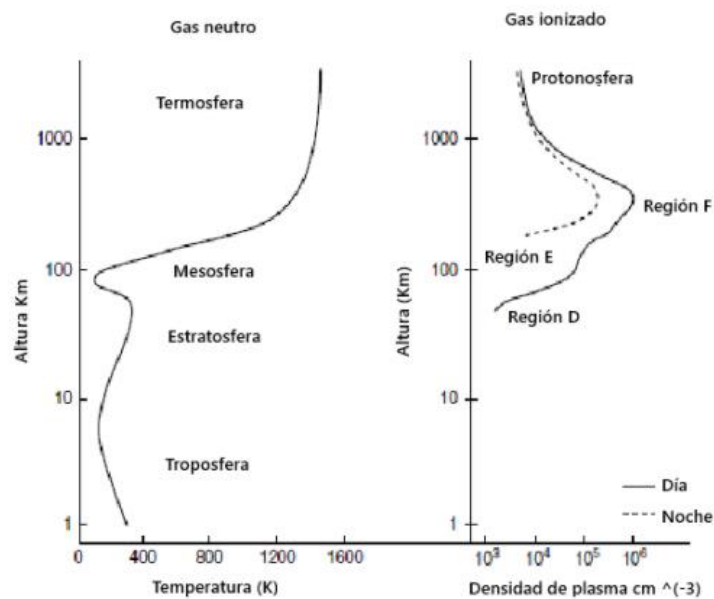


Figura 1.1: Perfil de temperatura de la atmósfera (panel izquierdo). Perfil de densidad de electrones en la ionosfera (panel derecho). Fuente: Hargreaves (1992).

La proporción fija en la composición de la atmósfera neutra (es decir 78% de N₂ y 20% de O₂) se mantiene solo hasta los ~ 100 km, gracias a procesos de mezcla turbulenta. Por encima de esta altura (la “turbopausa”), los constituyentes se separan según su masa (Rishbeth & Garriott, 1969; Hargreaves, 1992; Schunk & Nagy, 2009). A partir de allí las proporciones empiezan a cambiar, e incluso los componentes mayoritarios comienzan a ser otros. Como ejemplo, la Figura 1.2 muestra perfiles de concentración estimados con el modelo NRLMSISE-00 (Picone et al., 2002), desde la superficie de la Tierra hasta los 900 km de altura. Este es un modelo empírico atmosférico global de referencia (NRL por US Naval Research Laboratory, MSIS por Mass Spectrometer and Incoherent Scatter radar, E indica que el modelo se extiende desde la superficie hasta la exosfera, y 00 es el año de publicación). El modelo se puede correr online (<https://kauai.ccmc.gsfc.nasa.gov/instantrun/nrlmsis/>) y también está disponible en MatLab o Python.

El rango de altitud de 150 a 500 km se denomina región F de la ionosfera, y la densidad máxima en esta región se conoce como pico F. (Durante el día, la región F suele dividirse en F1 y F2 debido a la presencia de iones moleculares). Por debajo del pico, los iones NO⁺ y O⁺ adquieren mayor importancia, dominando el plasma por debajo de los 150 km

aproximadamente. El rango de altitud de 90 a 150 km se denomina región E, y la ionización por debajo de los 90 km se conoce, como era de esperar, como región D. Estos nombres tienen una historia curiosa. La región E recibió su nombre del campo eléctrico de la onda de radio reflejada por la capa "Heavyside" (el primer nombre que se le dio a la ionosfera). Las demás capas eran simplemente extensiones alfabéticas. Inicialmente se asumió que no existía plasma entre las capas. Desafortunadamente, muchos fenómenos han recibido su nombre del instrumento utilizado para medirlos o de algún otro parámetro poco claro.

Los componentes neutros que da el NRLMSISE-00 son N_2 , O_2 , N, O, Ar, He, H. En el panel izquierdo de la Figura 1.2 se graficó la concentración total acumulada comenzando desde la superficie. Ahí se ve claramente que hasta los 50 km de altura se acumuló casi la totalidad de la atmósfera. En el panel del centro se ve la concentración numérica de cada constituyente. En la homósfera vemos que las curvas son paralelas, es decir que las proporciones se mantienen. A partir de los 100 km ya cada curva se comporta diferente y vemos por ejemplo que entre los 250 y los 500 km, aproximadamente, predomina el oxígeno atómico y a partir de los 800 km el He y el H.

En el panel derecho se grafica la concentración porcentual, pero considerando, no la densidad numérica sino la de masa. Se observa claramente que por ejemplo el N_2 (curva azul) se mantienen en 78% desde la superficie hasta los 100 km y a partir de esa altura decrece. Se ve también, por ejemplo, que el componente mayoritario entre los 250 km y más de 500 km es el O.

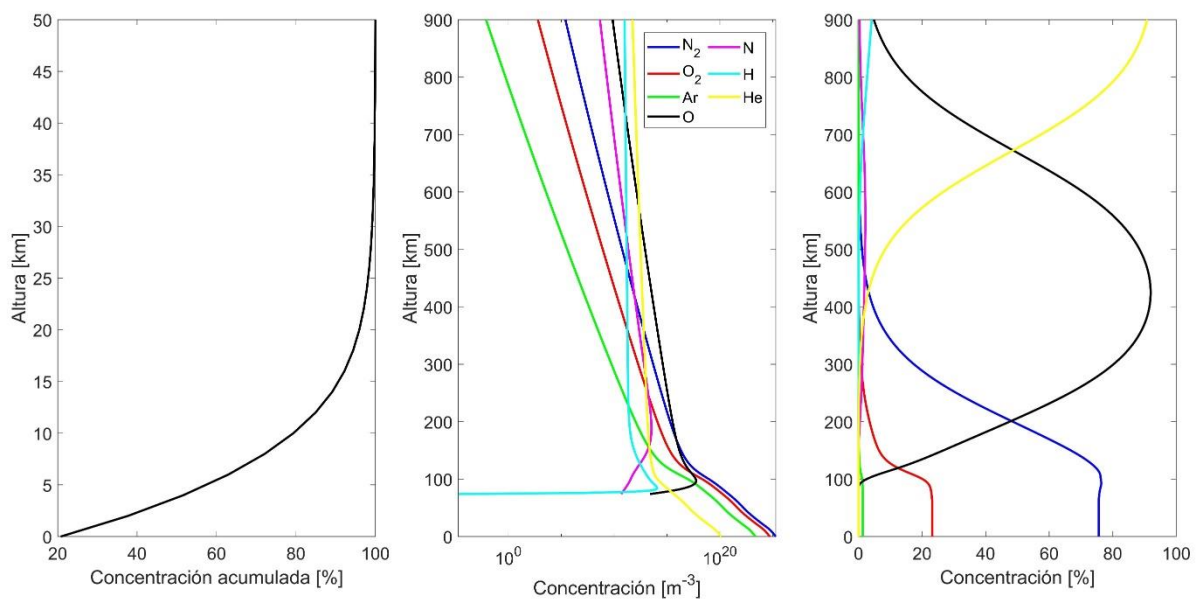


Figura 1.2 Concentración total acumulada en función de la altura (panel izquierdo). Concentración numérica de cada constituyente (panel del centro). Concentración porcentual estimada en base a la densidad de masa (panel derecho). Valores calculados con el modelo NRLMSISE-00 considerando F10.7=100, día del año: 15, $A_p=0$, latitud: 30°N, longitud: 0°.

1.3 Densidad electrónica en la ionosfera y procesos físicos que la afectan

El parámetro ionosférico más importante desde el punto de vista de la propagación de ondas de radio es la densidad electrónica. La densidad en un punto dado del espacio y del tiempo depende de tres procesos físicos: producción, pérdida y transporte. Estos tres procesos engloban fenómenos fotoquímicos y electrodinámicos complejos, influenciados por diversas condiciones externas de la atmósfera y la magnetosfera.

La fuente principal de ionización en la ionósfera es la radiación electromagnética solar en los rangos extremo ultravioleta y rayos X. Sin embargo, existen también otras fuentes relevantes, como partículas cargadas energéticas de origen solar y magnetosférico a altas latitudes. La tasa de ionización a distintas altitudes depende de la intensidad de la radiación solar, la longitud de onda y de la eficiencia de ionización de los gases neutros atmosféricos.

A medida que la radiación solar atraviesa la atmósfera, es absorbida progresivamente, por lo que su capacidad ionizante depende de la longitud del trayecto atmosférico, y por ende, del ángulo cenital solar, χ . La máxima tasa de ionización ocurre cuando el Sol está en el cenit ($\chi = 0$). A su vez χ varía también con la posición geográfica y día del año, lo que resulta en variaciones con la hora del día, estación del año y posición geográfica, que se describirán en el Capítulo 2.

La producción de iones libres y electrones por radiación solar se contrarresta con procesos de pérdida, principalmente la recombinación por colisión entre electrones e iones, y la captura de electrones por átomos y moléculas neutras. Haciendo algunos supuestos es posible describir de manera matemática simple la distribución vertical de la ionización, considerando el flujo de radiación solar ionizante, la distribución de constituyentes neutros y su eficiencia de absorción, junto con χ . Esto fue planteado por primera vez por Chapman (1931) en un modelo para una atmósfera isotérmica, de un solo constituyente, ionizada por radiación solar monocromática. Este modelo (modelo de Chapman) constituye una buena aproximación para las regiones E y F1, donde la fotoionización y la recombinación son dominantes, pero resulta menos adecuado en las regiones D y F2, donde intervienen otros procesos relevantes.

En la altura de la capa F2, el transporte de cargas adquiere importancia. Este transporte puede deberse a difusión, vientos neutros atmosféricos y derivas electromagnéticas. La difusión del plasma ocurre a lo largo de las líneas del campo magnético de la Tierra. Los vientos neutros tienen también su efecto, ya que impulsan la ionización de la zona diurna más cálida hacia la zona nocturna más fría, desplazando a las cargas mediante colisiones con los neutros. Este transporte se da también a lo largo de las líneas de campo magnético.

En la región F2 también es relevante la deriva electromagnética ($E \times B$), causada por la interacción de partículas cargadas con el campo magnético terrestre (B) y los campos eléctricos (E). Estos campos eléctricos son especialmente importantes cerca del ecuador, donde generan derivas verticales, y en las zonas aurorales y casquetes polares, donde producen derivas horizontales (Rishbeth & Garriott, 1969; Schunk & Nagy, 2009).

Durante el día, la combinación de los procesos físicos en la ionosfera da como resultado una estructura de capas como se muestra en la Figura 1.1. En cada una de ellas hay un máximo de densidad de electrones a una dada altura, por encima y por debajo de la cual esta densidad decrece. Se reconocen en general tres capas, o regiones, denominadas D, E y F. A su vez la región F se separa durante el día en dos capas las que se denominan F1 y F2 (Rishbeth & Garriott, 1969; Hargreaves, 1992; Schunk & Nagy, 2009). La capa F1, al igual que la región E, está bajo fuerte control solar y presenta características de tipo Chapman, aunque con una

dependencia diferente respecto al ángulo cenital solar. Durante la noche, la capa F1 desaparece y solo permanece la capa F2.

La capa F2 es la más prominente de la ionosfera, presenta las mayores densidades electrónicas y es la única capa que persiste de manera significativa durante la noche. El modelo de Chapman no la representa adecuadamente, ya que la capa F2 está fuertemente influenciada por los vientos neutros, la difusión ambipolar y otros efectos electrodinámicos. La relación entre la dirección del campo geomagnético y la de los vientos neutros y derivas electrodinámicas desempeña un papel fundamental en la estructura de la F2. Es precisamente la respuesta del plasma a estos procesos dinámicos, en presencia del campo geomagnético, la que da origen a las variaciones geográficas y temporales observadas en la capa F2.

La densidad máxima generalmente ocurre después del mediodía local, a veces hacia la tarde. La altura del máximo varía entre 250 y 350 km en latitudes medias, y entre 350 y 500 km en latitudes ecuatoriales.

En los primeros años de la investigación ionosférica, las desviaciones significativas respecto al comportamiento predicho por el modelo de Chapman fueron denominadas “anomalías”. Aunque esta descripción ya no es válida, ciertos términos se siguen usando aún hoy, como los siguientes:

- Anomalía estacional: tendencia observada de que NmF2 alrededor del mediodía local en invierno supere al valor de verano. Se atribuye a cambios estacionales en el cociente entre las densidades de O y N₂ de la atmósfera neutra.
- Anomalía ecuatorial: NmF2 durante el día es mayor a $\pm 15^\circ$ de latitud magnética hacia ambos lados del ecuador magnético que en el ecuador. Esto se debe a procesos electrodinámicos debido a la configuración particular del campo magnético de la Tierra combinado con los vientos neutros.

1.4 Propagación de ondas de radio en un medio ionizado

1.4.1 Teoría magneto-iónica y la ecuación de Appleton

La propagación de ondas de radio en un medio ionizado y magnetizado, como la ionósfera, se describe mediante la teoría magneto-iónica. Esta teoría fue desarrollada en las primeras décadas del siglo XX a partir de los experimentos pioneros de Marconi en comunicación de larga distancia, que culminaron con la transmisión transatlántica en 1901 (Hargreaves, 1992). Posteriormente, en 1902, Kennelly y Heaviside (de manera independiente) propusieron que las ondas de radio eran reflejadas por una capa electrificada en la atmósfera superior.

La teoría evolucionó gradualmente durante unos 30 años gracias a los aportes de múltiples investigadores. William Henry Eccles, por ejemplo, trató dicha capa como un reflector metálico en 1912, mientras que Joseph Larmor la consideró un dieléctrico en 1924. Este último enfoque es el más adecuado para describir la mayoría de las frecuencias de radio, ya que la absorción en el medio es relativamente pequeña y, en la mayoría de los casos, la onda retorna al suelo por refracción gradual, y no por reflexión abrupta (Hargreaves, 1992).

La formulación actualmente aceptada de la teoría se debe principalmente a los trabajos de E. V. Appleton y sus colaboradores entre 1927 y 1932 (Hargreaves, 1992). El resultado central de esta teoría es la conocida ecuación de Appleton–Hartree, que expresa el índice de refracción de

una onda de radio que se propaga en un medio ionizado sometido a un campo magnético, dado por

$$n^2 = 1 - \frac{X}{1 - jZ - \left[\frac{Y_T^2}{2(1 - X - jZ)} \right] \pm \left[\frac{Y_T^4}{4(1 - X - jZ)^2} + Y_L^2 \right]^{\frac{1}{2}}} \quad (1.1)$$

El índice de refracción n es complejo en el caso general ($n = \mu - jx$). Las cantidades adimensionales involucradas en la Ecuación (1.1) corresponden a:

$$X = \frac{\omega_N^2}{\omega^2}; \quad Y = \frac{\omega_B}{\omega}; \quad Y_L = \frac{\omega_L}{\omega}; \quad Y_T = \frac{\omega_T}{\omega}; \quad Z = \frac{\nu}{\omega}$$

Donde:

ω_N es la frecuencia angular de plasma

ω_B es la girofrecuencia del electrón

$\omega_L = \omega_B \cos(\theta)$ y $\omega_T = \omega_B \sin(\theta)$ son las componentes longitudinal y transversal de ω_B con respecto a la dirección de propagación

ν es la frecuencia de colisión de los electrones

ω es la frecuencia angular de la onda que se propaga

θ es el ángulo entre la dirección de propagación y el campo geomagnético

Si bien la Ecuación (1.1) es compleja, se puede simplificar bajo ciertas condiciones. Si se desprecia la absorción ($Z=0$) y los efectos del campo geomagnético ($Y_L = Y_T = 0$) la expresión se reduce a:

$$n^2 = 1 - X = 1 - \frac{\omega_N^2}{\omega^2} \quad (1.2)$$

Esta forma simplificada ya resulta útil para comprender los fundamentos de la propagación de ondas de radio en la ionosfera.

Cuando se considera el efecto del campo magnético, la ecuación de Appleton-Hartree se vuelve de doble valor, lo que implica que existen dos índices de refracción distintos, como ocurre en un cristal óptico birrefringente. Estos índices corresponden a las llamadas ondas características: el signo positivo en la ecuación define la onda ordinaria, mientras que el signo negativo define la onda extraordinaria. Las ondas de polarización arbitraria que ingresan a la ionósfera se descomponen en estas dos ondas características, que se propagan independientemente.

1.4.2 Reflexión ionosférica de ondas de alta frecuencia

La ionósfera desempeña un papel fundamental en las comunicaciones en la banda de alta frecuencia (HF), que abarca frecuencias entre 3 y 30 MHz, correspondientes a longitudes de onda entre 100 y 10 metros. Estas longitudes de onda son lo suficientemente cortas como para que el medio ionosférico no varíe significativamente a lo largo de unas pocas longitudes de onda. Por ello, la ionosfera puede modelarse como un conjunto de capas delgadas y uniformes, lo que constituye un medio de variación lenta. Este modelo permite comprender de forma sencilla por qué una señal de radio es reflejada por la ionósfera. Suponiendo una Tierra plana, se considera una señal de radio transmitida con un ángulo i_0 respecto del cenit. La ionósfera se

modela como una pila de capas delgadas, cada una de espesor mayor que una longitud de onda, con índices de refracción $n_1, n_2, n_3, \dots$. Aplicando la ley de Snell en cada frontera:

$$\begin{aligned}\sin i_0 &= n_1 \sin i_1 \\ n_1 \sin i_1 &= n_2 \sin i_2 \\ n_2 \sin i_2 &= n_3 \sin i_3 \\ &\vdots \\ n_{r-1} \sin i_{r-1} &= n_r\end{aligned}$$

$$\rightarrow \frac{\sin i_0}{n_r} = \text{constante}$$

A medida que aumenta la altura, la frecuencia del plasma ω_N también aumenta, y el índice de refracción n disminuye. Como consecuencia, el rayo se curva gradualmente hasta alcanzar una propagación horizontal, lo que ocurre cuando $n = \sin i_0$. Ese punto define el nivel de reflexión. Siempre que la ionósfera sea horizontalmente uniforme, el rayo retorna por una trayectoria simétrica. Utilizando la forma simplificada de la Ecuación (1.2) resulta:

$$\frac{\omega_N^2}{\omega^2} = 1 - n^2 = 1 - \sin^2 i_0 = \cos^2 i_0 \quad (1.3)$$

Por lo tanto, la condición de reflexión queda como:

$$\omega_N = \omega \cos i_0 \quad (1.4)$$

En el caso de incidencia vertical ($i_0 = 0$) se tiene que $\omega_N = \omega$ en el nivel de reflexión; es decir, la frecuencia del plasma es igual a la frecuencia de la onda de radio. Esto implica, según la Ecuación (1.3), que si $\omega_N > \omega$, el índice n se vuelve imaginario y la onda continúa propagándose sin reflejarse. La reflexión ocurre entonces cuando $n = 0$.

1.5 Deducción de la frecuencia de plasma

Supongamos una caja rectangular de plasma neutro compuesto por electrones y iones positivos, como se muestra en la Figura 1.3.

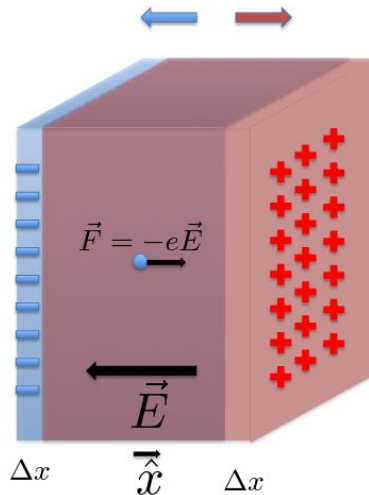


Figura 1.3 Electrones desplazado hacia la izquierda una distancia Δx , lo que genera un campo eléctrico E y una fuerza restauradora F .

Desplazamos el centro de masa de los electrones hacia la izquierda una distancia Δx . Esto genera una acumulación de electrones hacia la izquierda y de iones hacia la derecha, y en consecuencia un campo eléctrico E que apunta desde la "cara" o "placa" positiva de la caja hacia la negativa. Si se supone ahora que la distancia entre las placas positiva y negativa es muy pequeña en comparación con el área de las placas, entonces las condiciones de contorno están demasiado lejos de nuestros puntos de interés y podemos considerar esto como un condensador de placas paralelas ideal. Bajo esta suposición E resulta

$$E = \frac{\sigma}{\epsilon_0} = \frac{Q/A}{\epsilon_0} = \frac{(e N_e A \Delta x)/A}{\epsilon_0} = \frac{e N_e \Delta x}{\epsilon_0} \quad (1.5)$$

donde σ es la densidad superficial de carga por unidad de área de las "placas", Q es la carga total, y A el área, de manera que solo depende del espesor o separación entre las "placas" y la densidad de carga.

Si ahora suponemos un electrón en el medio de la caja, la fuerza que sentirá debido a E es

$$F = m a = -e E = \frac{e^2 N_e \Delta x}{\epsilon_0} \quad (1.6)$$

La aceleración a resulta entonces

$$a = \frac{e^2 N_e \Delta x}{m \epsilon_0} \quad (1.7)$$

La Ecuación 1.6 es simplemente la ecuación de un oscilador armónico simple con frecuencia dada por

$$\omega_N = \sqrt{\frac{e^2 N_e}{m \epsilon_0}} \quad (1.8)$$

que es la frecuencia de plasma de los electrones, donde el plasma tiene densidad de carga N_e .

De manera análoga puede calcularse la frecuencia de plasma de los iones reemplazando la masa m del electrón por la masa del ion en cuestión.

Si se aumenta N_e , para un mismo Δx , la fuerza será mayor y por lo tanto la oscilación será más rápida. También para un mismo campo E , la aceleración de los electrones será mayor que la de los iones porque la fuerza será la misma de manera que la aceleración depende inversamente de la masa de cada uno.

Reemplazando numéricamente en la Ecuación 1.8 los valores de e , m y ϵ_0 , y considerando la frecuencia f en vez de la frecuencia angular $\omega_N (=2\pi f)$ se obtiene para N_e

$$N_e = 1.24 \times 10^{10} f^2 \quad (1.9)$$

donde f está dada en MHz (que es la unidad usual en frecuencias de plasma en la ionosfera) y N_e en m^{-3} .

1.6 Técnicas de observación del entorno espacial de la Tierra

1.6.1 Importancia de las observaciones y clasificación de las técnicas de sensado

El desarrollo del conocimiento científico en el ámbito del entorno espacial terrestre depende de forma crítica de las observaciones. La ciencia se basa en datos del mundo real, y su avance está estrechamente vinculado a nuestra capacidad de registrar y analizar fenómenos mediante técnicas de medición cada vez más precisas. En regiones como la ionosfera o el espacio cercano a la Tierra, de difícil acceso directo, las observaciones instrumentales han sido esenciales para extender nuestra comprensión. De hecho, gran parte del crecimiento del conocimiento en geociencias durante las últimas décadas puede atribuirse al desarrollo de nuevos métodos observacionales, cuya efectividad depende no solo del instrumento utilizado, sino también de la forma en que se realiza la medición.

En general, las técnicas observacionales pueden clasificarse en tres categorías: sensado directo, indirecto y remoto:

- Sensado directo: implica la colocación de un instrumento dentro del medio a estudiar para medir directamente alguna propiedad física, como la temperatura o la densidad. Este enfoque suele ser específico y preciso, aunque puede alterar el medio y suele requerir vehículos (por ejemplo, satélites o globos) para su implementación, lo que incrementa significativamente los costos operativos.
- Sensado indirecto: también requiere colocar un objeto o sensor en el medio, pero este no mide directamente la magnitud de interés. En cambio, se infiere dicha magnitud observando el comportamiento del sensor desde una posición remota. Por ejemplo, es posible estimar la densidad o la velocidad del viento mediante el análisis del movimiento del vehículo en el que se encuentra el sensor. Aunque menos invasivo, este enfoque también requiere equipamiento móvil y puede presentar limitaciones en cuanto a resolución o precisión.
- Sensado remoto: se caracteriza por no requerir la inserción física de instrumentos en el medio. Las propiedades del entorno se deducen a partir del análisis de ondas (electromagnéticas, por ejemplo) que lo atraviesan o que son emitidas desde él. Este sensado puede ser activo, si la señal es generada artificialmente y enviada hacia el medio (como en un radar), o pasivo, cuando se observan señales generadas de forma natural (como la radiación de una aurora). Esta última categoría, el sensado remoto pasivo, es particularmente relevante en estudios ionosféricos.

La diversidad de técnicas de sensado remoto tienen ventajas en términos generales por su eficiencia operativa, continuidad temporal y bajo requerimiento de mantenimiento del instrumental. Sin embargo, la interpretación de los datos puede presentar desafíos, ya que las señales registradas pueden haber sido afectadas por múltiples procesos durante su propagación. A pesar de ello, el sensado remoto representa una herramienta fundamental para el estudio de regiones inaccesibles como la ionosfera.

1.6.2 Sondeo ionosférico por ondas de radio

Las técnicas de detección remota por ondas de radio constituyen una de las principales herramientas para el estudio de la atmósfera superior. Basadas en la propagación de ondas de radio, han permitido obtener una enorme cantidad de información.

Estas técnicas pueden clasificarse en tres grupos:

a- La onda de radio puede ser totalmente reflejada dentro del medio

b- Puede atravesar el medio y emerger alterada

c- La mayor parte de la energía puede atravesar el medio, pero una fracción se dispersa o refleja parcialmente.

Los grupos (a) y (c) requieren un transmisor y un receptor ubicados ambos por debajo o ambos por encima de la ionósfera, mientras que el grupo (b) generalmente necesita una fuente o un receptor por encima de la misma. Las técnicas del grupo (c) requieren mayor sensibilidad debido a que las señales reflejadas son más débiles.

Una de las técnicas más antiguas, y aún la más usada, es el sondeo ionosférico por medio de la ionosonda que se describe a continuación. El parámetro ionosférico utilizado en el análisis de llenado de datos faltantes de esta Tesis, foF2, es precisamente un parámetro medido por ionosondas, por lo que se describirá más en detalle su funcionamiento y características.

1.6.3 Ionosonda

Como se mencionó en la Sección 1.3.2, cuando una onda electromagnética viaja verticalmente y penetra en la ionosfera llega una altura a la cual el índice de refracción n se hace cero, y la onda se refleja. En esto se basa la ionosonda, o sondador ionosférico, para obtener el perfil de densidad de electrones de esta región atmosférica: transmite un pulso de radio verticalmente y mide el tiempo que tarda la señal en regresar. Este es el instrumento de sensado remoto de la ionosfera más antiguo y aún el más usado.

La ionosonda trabaja en un rango de frecuencias comprendido entre 1 y 30 Mhz. Consiste en un transmisor y un receptor de pulsos como se muestra esquemáticamente en la Figura 1.3. El tiempo transcurrido entre la emisión y la recepción representa el tiempo de propagación hacia y desde la ionosfera dado por

$$\Delta t = 2 \frac{h'}{c} \quad (1.10)$$

donde h' es la altura aparente o “altura virtual” donde se refleja el pulso. El nombre de altura aparente se debe a que ésta se calcula considerando que el pulso viaja con velocidad c cuando en realidad no es así. El medio que atraviesa la onda cuando ingresa a la ionosfera tiene un índice de refracción, n , que depende la densidad de plasma (Ecuaciones 1.1 y 1.2) y que varía a través del recorrido real h produciendo un retardo en el tiempo de "viaje" de la onda. Es decir que h' resulta mayor que la altura real de reflexión.

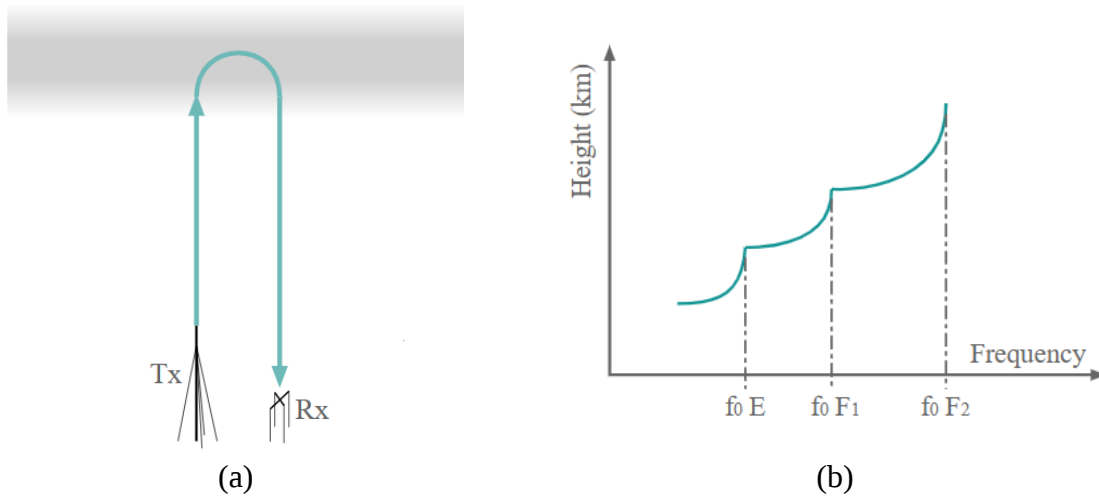


Figura 1.3 (a) Esquema de una ionosonda: Tx: transmisor, Rx: receptor, la zona sombreada de gris corresponde a la ionosfera, y la flecha verde corresponde a la señal transmitida por Tx que penetra en la ionosfera donde comienza a refractarse hasta reflejarse y regresar a la superficie de la Tierra, a Rx. (b) Diagrama idealizado de un ionograma: curva de frecuencia en función de la altura virtual de reflexión. Fuente: <https://www.electronics-notes.com/articles/antennas-propagation/ionospheric/ionosonde-ionogram.php>

El registro del sondador consiste en una gráfica de la altura aparente de reflexión h' versus la frecuencia, y se denomina ionograma o curva $h'(f)$ como se observa de manera esquemática en la Figura 1.3.

En el pico de una capa ionosférica, donde N_e es máxima, a partir de una cierta frecuencia, f_o , la señal penetra dicha capa. f_o , la mínima frecuencia de penetración de la capa, se denomina frecuencia crítica. La máxima densidad de electrones de cada capa, E, F1 y F2, se obtiene entonces reemplazando f en la Ecuación 1.10 por la frecuencia crítica de cada capa: f_oE , f_oF1 y f_oF2 respectivamente. Resulta así para la máxima densidad de electrones de las regiones E, F1 y F2 de la ionosfera

$$NmE = 1.24 \times 10^{10} f_oE^2 \quad (1.11)$$

$$NmF1 = 1.24 \times 10^{10} f_oF1^2 \quad (1.12)$$

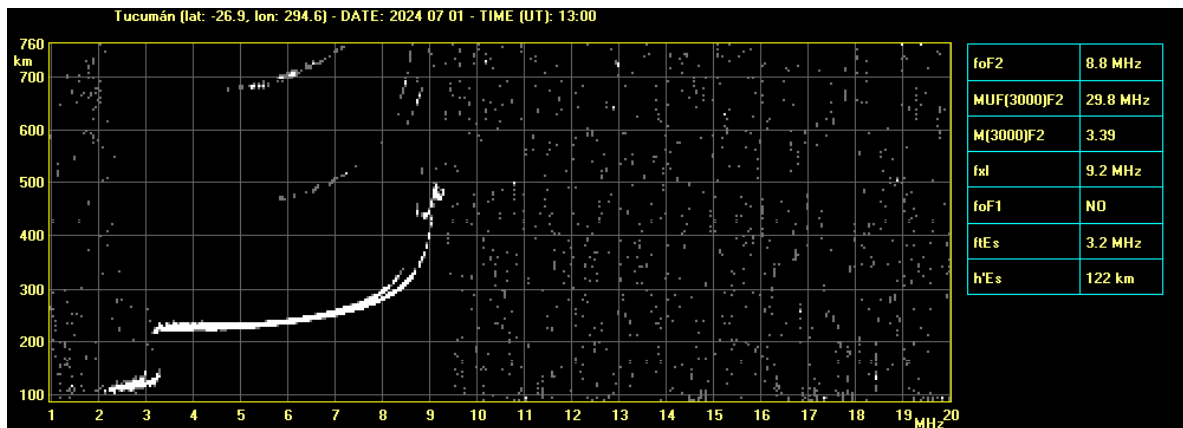
$$NmF2 = 1.24 \times 10^{10} f_oF2^2 \quad (1.13)$$

La determinación de la altura real de reflexión requiere la inversión de la ecuación integral

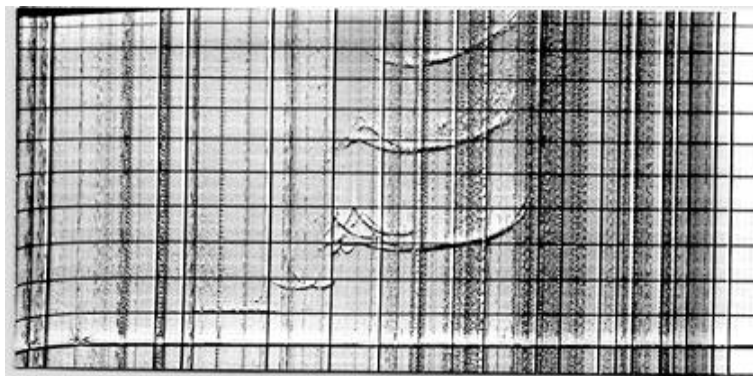
$$h' = \int_0^h \mu'(f, N_e) dh \quad (1.14)$$

donde μ' , el índice de refracción de grupo del medio que atraviesa la onda, dada por la inversa del índice de refracción (Ecuación 1.1 y 1.2), lo cual es complejo y muy laborioso.

En la Figura 1.4 se muestra un ionograma obtenido con una ionosonda y otro con una digisonda.



(a)



(b)

Figura 1.4. Ionograma obtenido con (a) la digisonda del Observatorio de Tucumán y (b) con la ionosonda que funcionó entre los años 1957-1987 en la Estación Ionosférica de Tucumán, en el Parque 9 de Julio. (<http://ionos.ingv.it/tucuman/ionograms/>)

A partir de 1930 la red de estaciones de sondeo vertical de la ionosfera mediante ionosondas comenzó a expandirse considerablemente (Zolessi & Cander, 2014). La máxima expansión de esta red se dio en el Año Geofísico Internacional (IGY - International Geophysical Year), julio de 1957 a diciembre de 1958. En ese año se instaló también una ionosonda en Tucumán, en un predio del Parque 9 de Julio, manejada por un grupo del Instituto de Ingeniería Eléctrica (hoy Departamento de Ingeniería Eléctrica) de la FACET, que funcionó hasta 1987.

La digisonda (acrónimo de Digital Ionospheric Goniometric ionoSONDE) es una ionosonda de última generación. A diferencia de los sistemas analógicos más antiguos, que imprimían los ionogramas en papel (como se ve en la Figura 1.5b) y requerían la intervención manual de un operador para extraer parámetros como foF2, foE o la altura virtual, la digisonda permite la digitalización automática de los ionogramas, lo que facilita su almacenamiento, transmisión y análisis en tiempo real (Reinisch et al., 2009).

Uno de los componentes más relevantes del sistema es el software ARTIST (Automatic Real-Time Ionogram Scaler with True height analysis), encargado de identificar las trazas principales (ordinaria y extraordinaria) y realizar la conversión de alturas virtuales en alturas reales, obteniendo así un perfil de densidad electrónica consistente con la física de propagación de ondas en el plasma ionosférico. ARTIST constituye un avance significativo, ya que reemplaza

el análisis manual que históricamente realizaban los operadores, proporcionando resultados más rápidos y estandarizados (Reinisch et al., 1983).

El término 4D en la denominación Digisonde 4D (DPS4D) hace referencia a que el sistema no solo registra las dimensiones clásicas del ionograma (frecuencia y altura virtual), sino que también incorpora la variabilidad temporal y espacial mediante la detección direccional de los ecos. Esto permite estudiar desplazamientos de plasma (drifts) y estructuras dinámicas en la ionosfera, añadiendo así una dimensión adicional de información al análisis ionosférico (Reinisch et al., 2009).

En cuanto a la configuración instrumental, las digisondas utilizan arreglos de antenas transmisoras y receptoras, dispuestas de manera tal que se optimiza la resolución angular y el campo de visión del instrumento. Parámetros como la separación máxima de antenas, la desviación relativa y la rotación del arreglo se ajustan en función de las condiciones locales y de los objetivos científicos de la instalación.

Actualmente, la mayoría de las digisondas forman parte de la red internacional GIRO (Global Ionosphere Radio Observatory), que proporciona ionogramas estandarizados con una resolución temporal típica de 15 minutos. Esto convierte a la digisonda en una herramienta fundamental tanto para el monitoreo operativo del clima espacial como para la investigación de la dinámica ionosférica en todas sus escalas.

La red de ionosondas provee la base de datos más extensa en el tiempo disponible y organizada siguiendo reglas standard de mediciones y continúa siendo la herramienta básica en la investigación ionosférica. El monitoreo ionosférico es importante en el contexto del Clima Espacial, con aplicaciones en comunicaciones de radio, navegación y predicción de órbitas.

En particular, es esta base de datos la única hasta el momento que posee las características necesarias para estudiar tendencias y variaciones de largo plazo ya que un requisito fundamental para obtener valores con significancia estadística, es que la serie tenga una extensión temporal de dos ciclos solares completos (Lastovicka et al., 2006), y aún más.

Capítulo 2. Variabilidad temporal del parámetro ionosférico foF2 y sus principales forzantes

El conocimiento previo de la variabilidad en distintas escalas, tanto temporal como espacial, resulta fundamental, no solo para interpretar la dinámica del sistema de estudio (la ionosfera en este caso), sino que constituye una base necesaria al aplicar métodos de ML en el completado de series con datos faltantes, eje central de esta Tesis. Con respecto a la variación temporal, que es la de interés en este trabajo, la ionosfera, y en particular la capa F2, varía en escalas temporales que van desde minutos hasta años. Esta variabilidad resulta de una amplia gama de procesos físicos de origen tanto terrestre como solar. Comprender estos procesos y sus interacciones para predecir estas variabilidades es extremadamente complejo y sigue siendo un gran desafío. En este capítulo se analiza, de manera descriptiva, la variabilidad regular del parámetro ionosférico foF2, como lo son las variaciones diaria y estacional, las asociadas a la actividad solar y geomagnética. Se describen además los principales forzantes que definen algunos de los principales patrones de variación temporal.

2.1 Variabilidad espacial de la capa F2

Si bien en este trabajo no se hace un interpolado espacial, y tampoco se analiza el método de interpolación de datos en la serie temporal de una dada estación utilizando datos de estaciones próximas, se presenta de manera cualitativa y breve algunos aspectos de la variabilidad espacial de foF2.

Los principales forzantes que generan el patrón espacial de variabilidad de foF2 son dos: la intensidad de la radiación solar que, para un instante dado, varía con la latitud geográfica, y el campo magnético de la Tierra que, al ser prácticamente dipolar, varía con la latitud geomagnética. El primer forzante afecta fundamentalmente la producción de ionización, y el segundo a los procesos de transporte, ya que en esta región se tienen cargas libres en movimiento, con una frecuencia de colisión menos dominante que la girofrecuencia debido a la presencia del campo geomagnético.

La radiación solar recibida disminuye al aumentar la latitud, ya que a altitudes más bajas (cerca del ecuador) los rayos del sol inciden más directamente, concentrando la energía en un área menor, mientras que a latitudes más altas los rayos se distribuyen sobre un área mayor, lo que resulta en una menor intensidad de energía por unidad de superficie. Esto haría esperar, durante horas del día, mayor densidad de electrones y por lo tanto mayor foF2 en el ecuador geográfico y una disminución hacia latitudes menores. Sin embargo, esta variabilidad espacial está regulada por procesos dinámicos que responden a la configuración del campo magnético de la Tierra y que hacen que los máximos de foF2 ocurran a $\pm 15^\circ$ de latitud geomagnética, a ambos lados del ecuador magnético, con una depresión en el valor de foF2 a lo largo de este ecuador. Este proceso da lugar a la anomalía ecuatorial que se mencionó en la Sección 1.3. La Figura 2.1 muestra, como ejemplo, la distribución espacial de foF2 para las 12 LT, en el mes de marzo de 2000, estimada con el modelo IRI (International Reference Ionosphere) (Bilitza et al., 2022) donde se ve claramente la depresión en foF2 en la región sobre el ecuador magnético, que, en el presente, se ubica próximo al ecuador geográfico en casi todas las longitudes. El modelo IRI se puede correr online (<https://kauai.ccmc.gsfc.nasa.gov/instantrun/iri/>) y también está disponible en MatLab o Python.

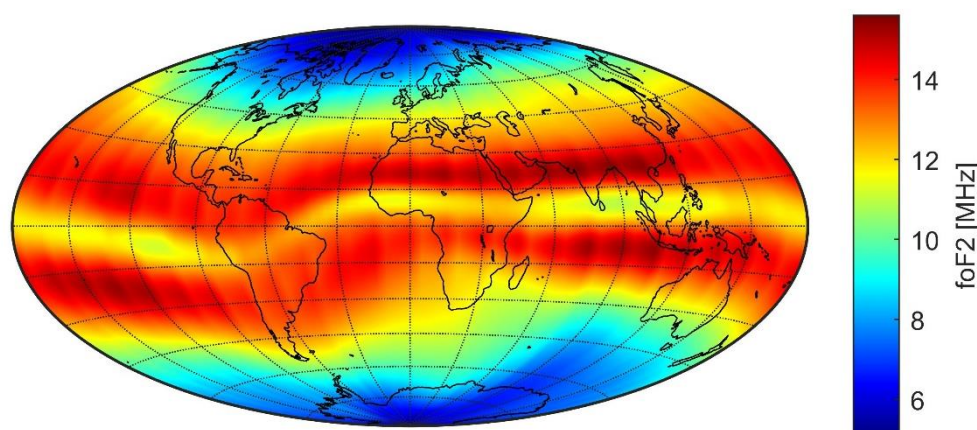


Figura 2.1 Frecuencia crítica de la capa F2, foF2, estimada con IRI-2020, considerando el mes de marzo de 2000, 12 LT, y desactivando el modelo de tormenta ("storm off").

Si bien no entraremos en detalles sobre estos patrones de variabilidad espacial de foF2, es importante tenerlos en cuenta ya que dan una idea de la longitud de la correlación espacial que es fundamental considerar para interpolaciones espaciales en donde se utilizan datos de otras estaciones con registros más completos para interpolar datos faltantes en el punto de interés.

2.2 Variación diurna

La variación diurna del parámetro foF2 refleja el ciclo diario de ionización y recombinación.

Sus características principales son: el mínimo nocturno se registra generalmente entre la medianoche y el amanecer (aproximadamente 03-06 LT). Durante la noche, la ausencia de radiación solar ultravioleta provoca una disminución progresiva de la densidad electrónica por recombinación, alcanzando su valor más bajo justo antes del amanecer. Por otro lado, el máximo diurno ocurre típicamente entre las 12 y las 16 LT, con un pico más frecuente alrededor de las 14-15 LT en latitudes medias. Este máximo responde principalmente a la intensa ionización producida por la radiación solar EUV que maximiza cerca del mediodía.

Las causas principales son la producción de ionización dominada por la radiación solar EUV que ioniza el oxígeno atómico (O) en la capa F2, siguiendo aproximadamente el ángulo cenital solar. Las pérdidas se deben principalmente a la recombinación con O₂ y N₂, junto con el transporte vertical inducido por vientos neutros y campos eléctricos

Esta variación diurna es fuertemente modulada por la actividad solar, la estación del año y las condiciones geomagnéticas. En condiciones de calma solar, el rango típico en latitudes medias va de aproximadamente 4-6 MHz durante la noche a 8-12 MHz en el mediodía.

2.3 Variación estacional

La variación estacional del parámetro foF2 muestra patrones anuales y semi-anuales que responden principalmente a cambios en la composición atmosférica, el ángulo cenital solar y la

circulación termosférica. Sus características principales son: en latitudes medias, se observa comúnmente la anomalía de invierno, donde los valores de foF2 son más altos durante el invierno que en verano, a pesar de recibir menos radiación solar. En latitudes bajas ecuatoriales, predomina una variación semianual con máximos pronunciados en los equinoccios.

2.4 Variación con la actividad solar

La variación de la actividad solar se debe básicamente a la variación del campo magnético del Sol generado por la acción de un dínamo que opera en su interior. Todos los aspectos del Sol, aunque con diferente intensidad, varían con la actividad solar cuya variabilidad más importante es la del ciclo de ~10-11 años, cuyo periodo y amplitud no son constantes, y se denomina ciclo de actividad solar. Hay que aclarar que la actividad solar presenta también otros ciclos en diferentes escalas de tiempo.

El parámetro foF2 responde a la actividad solar dado que responde al rango EUV de radiación. Este rango varía con la actividad solar en un ~100% entre un mínimo y un máximo, mientras que la irradiancia total solar (TSI) (que corresponde a la integral sobre todas las longitudes de onda del espectro solar) varía en menos del 0.1%. Es por eso que la ionosfera es sensible a la variación cuasi-decadal de la actividad solar y el clima de la troposfera, que es sensible al TSI, no. En realidad, es sensible al visible y al infrarrojo de la radiación solar que constituyen la mayor parte del espectro solar. Tener en cuenta que el rango EUV solar corresponde a menos del 0.1% del espectro total.

Para hacer una estimación más precisa de estos valores se utilizó la serie "Yearly SSI calculated using NNL1 solar irradiance model" que incluye la irradiancia total integrada TSI, NNLSSI1_v03r00, del "Laboratory for Atmospheric and Space Physics" (LASP). Este Registro de Datos Climáticos (CDR, por sus siglas en inglés) contiene la irradiancia solar espectral (SSI) y el TSI como función del tiempo y de la longitud de onda, creada mediante el modelo NASA NOAA LASP (NNL). La SSI representa el flujo energético dependiente de la longitud de onda que llega a la parte superior de la atmósfera terrestre, a una distancia estándar de una Unidad Astronómica del Sol. Sus unidades son ($\text{W/m}^2\text{nm}$), y el TSI, como ya se mencionó, corresponde a la integral sobre todas las longitudes de onda del espectro solar también en el tope de la atmósfera en unidades de (W/m^2). El conjunto de datos de SSI abarca el rango espectral que cubre desde aproximadamente 0 nm hasta 200000 nm. El conjunto de datos fue desarrollado por Odele Coddington, Judith Lean, Chris Lindholm y Peter Pilewskie del Laboratorio de Ciencias Atmosféricas y del Espacio (LASP), Universidad de Colorado.

El archivo de datos está en formato netCDF-4 (ssi_v03r00_yearly_s1610_e2024_c20250221.nc) disponible en <https://www.ncei.noaa.gov/data/solar-spectral-irradiance/access/standard-resolution/yearly/>.

La Figura 2.2 muestra la variación en el período 1960-2023 del rango EUV y del TSI, a partir de estos datos obtenidos del National Centers for Environmental Information (NCEI, <https://www.ncei.noaa.gov/>) que fueron reconstruidos desde el año 1610, donde se ve claramente que ambos varían con la actividad solar. Sin embargo, en términos porcentuales, utilizando estos valores, la variación entre un máximo y un mínimo de actividad solar (por ejemplo, entre 2002 y 1996), en el rango EUV (entre 10 y 110 nm) da de ~35% mientras que en el caso del TSI la variación es de un ~0.04%. Es decir, que el EUV varía 1000 veces más que el TSI. Esto explica en parte por qué la troposfera no es sensible a la variabilidad decadal del ciclo solar, mientras que la ionosfera y la termosfera sí lo son. En la misma Figura 2.2, se

muestra la variación porcentual entre un mínimo y un máximo de actividad solar, para diferentes longitudes de onda partiendo de 0.5 nm hasta 350 nm, a partir de donde la sensibilidad con el ciclo solar cae prácticamente a cero. Se observa incluso que para ciertas longitudes de onda, dentro del rango EUV, esta variación supera al promedio de 35%, llegando a casi 100%.

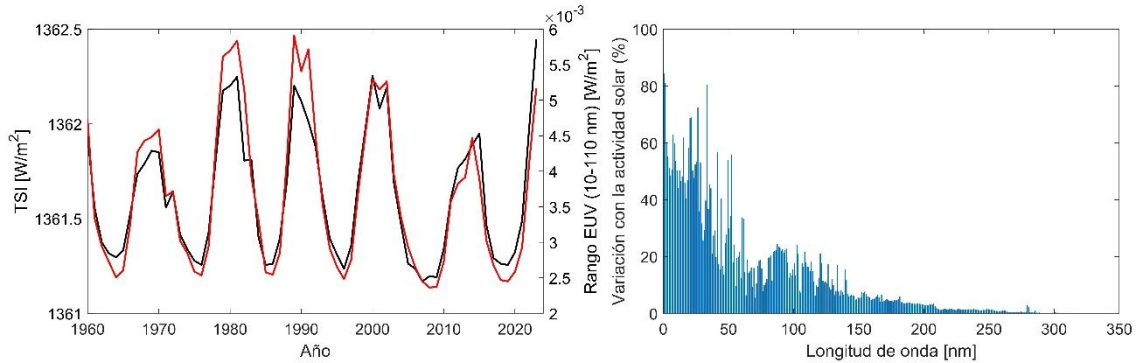


Figura 2.2 Irradiancia total solar (TSI, Total Solar Irradiance) [W/m^2] (línea negra) e integral del rango EUV (10 a 110 nm) [W/m^2], en el período 1960-2023 (basado en una reconstrucción ya que hay datos de EUV y de TSI satelitales desde 1978) (panel izquierdo). Variación porcentual entre un máximo y un mínimo de actividad solar respecto al nivel promedio de radiación, en función de la longitud de onda.

Como el rango EUV solar no llega a la superficie de la Tierra, se cuenta con mediciones recién desde 1978 cuando comienza la era satelital. A esto se suma el problema de la rápida degradación de los instrumentos a bordo de los satélites que introducen sesgos en las mediciones. Si bien esto es tenido en cuenta a la hora de homogeneizar y concatenar mediciones para obtener series largas de tiempo, algunas incertidumbres persisten.

En general, se utilizan proxies solares para determinar la variabilidad del EUV solar que responde a la actividad solar.

2.5 Proxies de la actividad solar

Existen varios proxies de la actividad solar, basados en índices que se originan en distintas regiones de la atmósfera solar. Si bien difieren en la variabilidad diaria, todos ellos presentan una variación decadal similar, aunque no idéntica. Nosotros solo consideraremos cuatro: número de manchas solares SN, MgII, F10.7, F30. Las manchas solares se originan en la fotosfera del Sol, el MgII en la alta cromosfera, y los flujos en ondas de radio F10.7 y F30 en la alta cromosfera y en la corona (Elias et al., 2023). La Figura 2.3 muestra las series de los valores promedios mensuales de estos proxies en el período analizado en este trabajo. Se observa claramente que la variación interanual es similar para todos, sobresaliendo la variación asociada al ciclo de actividad solar de ~ 11 años.

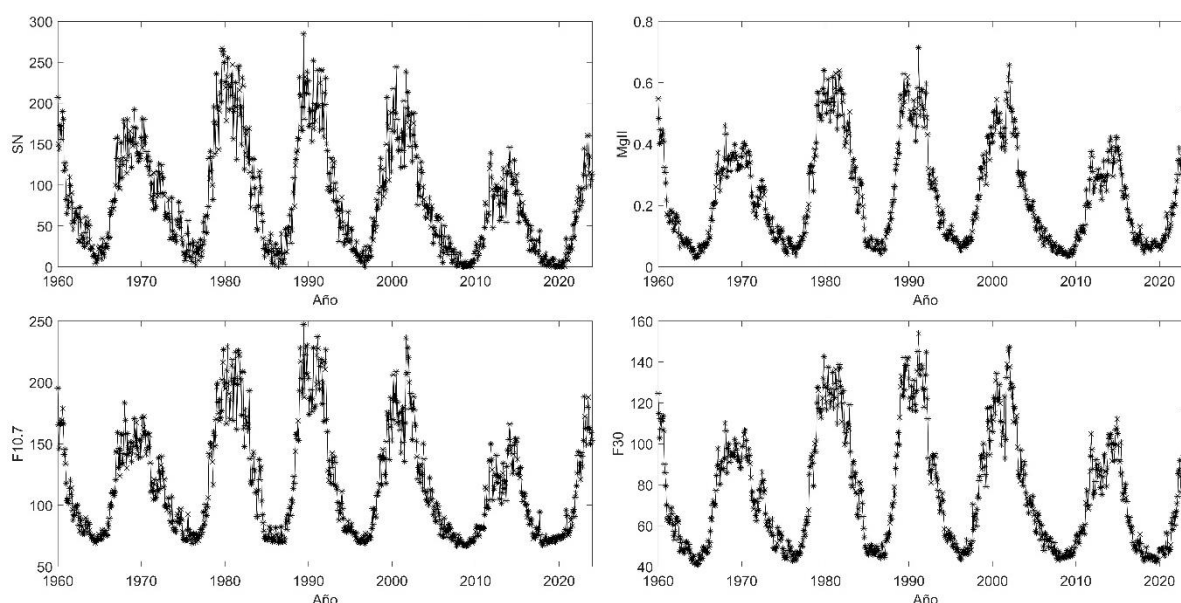


Figura 2.3 Series del promedio mensual de los proxies de EUV solar utilizados en esta Tesis: SN, MgII, F10.7 y F30 en el período enero 1960-diciembre 2023.

2.5.1 Número de manchas solares, SN

El número de manchas solares (Sunspot Number, SN) es uno de los indicadores más antiguos y más usados de la actividad solar. Constituye la base para el estudio del ciclo solar de aproximadamente 11 años y de la influencia a largo plazo del Sol sobre los distintos sistemas y procesos de la Tierra (Clette et al., 2014).

El número relativo de manchas solares fue introducido por Rudolf Wolf en 1849 y se calcula mediante la expresión $R=k(10N_g+N_s)$ donde N_g es el número de grupos de manchas, N_s el número total de manchas individuales, y k un coeficiente que corrige las diferencias entre observadores e instrumentos. Este índice no tiene una unidad física, por lo que Wolf lo denominó “número relativo”. Precisamente fue Wolf quien inició la serie histórica de lo que se denominó Número Internacional de Manchas Solares, continuada luego por el Observatorio de Zúrich. Esta serie se denominaba usualmente R_z . Actualmente es mantenida por el World Data Center SILSO en Bruselas y es denominada SN (<https://www.sidc.be/SILSO/datafiles>). Esta serie ha sido objeto de extensas revisiones debido a inconsistencias y sesgos históricos. La recalibración más reciente (Clette et al., 2014) integró observaciones adicionales y métodos modernos, dando origen al “Sunspot Number Version 2.0”, una versión actualizada y homogénea.

SN sigue siendo un parámetro esencial en Física Solar y ciencias del Clima Espacial. Su correlación con otros índices, como el flujo de radio solar F10.7 o el flujo de radiación EUV solar, lo convierte en un proxy clave de la radiación solar incidente en la atmósfera superior, y por lo tanto, en una variable de referencia para estudios ionosféricos como los que motivan este trabajo.

2.5.2 Índice MgII

El índice Mg II se basa en la doble línea espectral del magnesio ionizado (Mg II h y k), centradas en torno a 280 nm, que se originan en la región superior de la cromosfera solar. Este índice cuantifica la relación entre la intensidad del núcleo (core) y las alas (wings) de dichas líneas, conocidas como "core-to-wing ratio".

Las variaciones del MgII reflejan cambios en la radiación ultravioleta emitida por el Sol, ya que el núcleo de las líneas se forma en capas más altas y activas de la atmósfera solar, sensibles a la actividad magnética, mientras que las alas se originan en regiones más profundas y menos variables. Este índice se considera un excelente proxy de la radiación ultravioleta extrema (EUV) y se utiliza ampliamente para estudios de interacción Sol-Tierra y para la modelización de la ionosfera y la termosfera.

El MgII se obtiene de mediciones satelitales por lo que su registro comienza no antes de 1978. La serie de MgII se puede obtener de la Universidad de Bremen (<https://www.iup.uni-bremen.de/UVSAT/data/>), quienes derivaron una serie extensa concatenando las mediciones hechas por distintos satélites como se observa en la Figura 2.4.

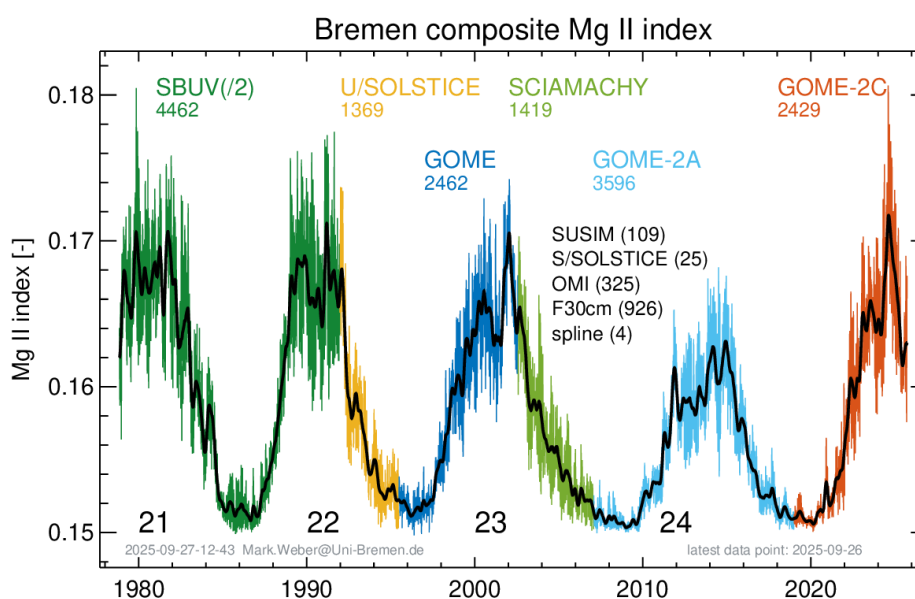


Figura 2.4 Índice combinado de Mg II (“Bremen composite MgII”). Los valores indican la cantidad de días aportados por cada satélite que contribuye al índice combinado (o compuesto). La curva negra muestra la serie temporal suavizada dos veces con un promedio móvil de 55 días. Fuente: <https://www.iup.uni-bremen.de/UVSAT/data/>

En este trabajo de Tesis, donde se considera el período de tiempo 1960-2023, se utilizó la serie extendida provista por la Universidad de Bremen (<https://www.iup.uni-bremen.de/UVSAT/data/>), Alemania, que extendió la serie original llevándola a 1947 utilizando ajustes lineales con F10.7 y F30. Esta serie está normalizada entre 0 y 1.

2.5.3 Flujo en onda de radio F10.7

Cabe destacar que este proxy solar despertó un especial interés durante el desarrollo de este trabajo.

El índice F10.7, es una medida del flujo de radio solar a 10.7 cm (2800 MHz) (Tapping, 2013), constituye uno de los registros más antiguos y confiables de la actividad solar. Su origen se remonta a mediados del siglo XX, en el contexto de la posguerra, cuando los avances en tecnología radar permitieron reutilizar equipamiento militar con fines científicos.

En Canadá, el físico Arthur Covington, junto a su equipo en el Consejo Nacional de Investigación en Ottawa, construyó el primer radiotelescopio del país reutilizando componentes de radar excedentes. Por razones técnicas, el instrumento operaba a una frecuencia de 2800 MHz (longitud de onda de 10.7 cm), lo cual determinó el nacimiento del índice. Esta longitud de onda, corresponde dentro del rango de ondas de radio, a las microondas. Estas últimas cubren el espectro de longitudes de onda entre 1 y 30 cm.

Las primeras observaciones se realizaron en 1946, y a partir de 1947 comenzaron las mediciones sistemáticas. Covington descubrió que las emisiones solares en esa banda variaban con el tiempo y estaban correlacionadas con la presencia de regiones activas como las manchas solares. Este hallazgo consolidó al F10.7 como un proxy confiable de la actividad solar.

Desde entonces, la medición continua del índice ha sido mantenida por Canadá, actualmente desde el National Research Council de Canadá, con la participación del Natural Resources Canada y la Canadian Space Agency (<https://spaceweather.gc.ca/forecast-prevision/solar-solaire/solarflux/sx-en.php>) acumulando así más de siete décadas de datos homogéneos, cruciales para estudios de clima espacial, heliofísica e ionosfera.

El índice F10.7 se expresa en unidades de flujo solar (sfu), donde $1\text{sfu} = 10^{-22}\text{Wm}^{-2}\text{s}$. Esta emisión de radio proviene de las capas altas de la cromosfera y de la parte baja de la corona solar. Tiene dos orígenes principales: la emisión térmica de Bremsstrahlung, producida cuando los electrones emiten radiación al ser desviados por otras partículas cargadas, y la emisión de giro-resonancia (o giroemisión), generada cuando los electrones emiten energía al cambiar de dirección al girar alrededor de las líneas del campo magnético solar.

El F10.7 es un índice especialmente sensible a la actividad magnética solar, ya que sus variaciones reflejan directamente la aparición y evolución de regiones activas. Tiene además varias ventajas técnicas: se mide desde tierra en cualquier condición meteorológica, posee una larga serie histórica continua y su comportamiento correlaciona bien con otros parámetros solares y geofísicos.

Por ello, el F10.7 es extensamente utilizado como proxy en modelos ionosféricos y atmosféricos, como predictor en modelos de propagación de ondas de radio, y como indicador global del nivel de actividad solar. Incluso ha permitido detectar cambios sutiles en el comportamiento del Sol, como la disminución del contraste de las manchas solares en los últimos ciclos, lo que sugiere una debilitación de los campos magnéticos.

En resumen, el índice F10.7 constituye un “termómetro” integral del Sol, con aplicaciones tanto científicas como tecnológicas, y su continuidad y calibración cuidadosa lo convierten en un pilar de la climatología espacial moderna.

2.5.4 Flujo en onda de radio F30

El índice F30 (en $\text{sfu}=10^{-22} \text{ Ws/m}^2$) corresponde a la emisión de radio del Sol en 30 cm (1 GHz). La fuente original de estos datos es el National Astronomical Observatory de Japón (https://solar.nro.nao.ac.jp/norp/data/daily/TYKW-NoRP_dailyflux.txt). Pero esta serie de datos tiene algunos datos faltantes para los meses de marzo, abril y mayo de 1994, que se completaron con la base de datos disponible en <https://spaceweather.cls.fr/services/radioflux/>. También se la puede obtener de https://lasp.colorado.edu/lisird/data/cls_radio_flux_f30

El flujo de radiación correspondiente al índice F30 es medido por el radio-polarímetro Toyokawa/Nobeyama, en Japón, que monitorea varias frecuencias en el rango de ondas de radio (entre ellas: 15 cm, 8 cm y 3.2 cm). Estas observaciones comenzaron en noviembre de 1951 en Tokoyama hasta 1994 cuando las antenas se trasladaron a Nobeyama (por eso el gap de los datos mencionados en el párrafo anterior) (Dudok de Wit et al., 2014). Cabe mencionar que todas estas longitudes de ondas, al igual que en el caso de F10.7, caen en el rango, además del de ondas de radio, en el de las microondas.

Este índice cobró importancia a partir de los trabajos de Lastovicka & Buresova (2023) y Lastovicka (2024), quienes indican que el proxy óptimo para detectar tendencias de largo plazo en foF2 es precisamente F30.

2.6 Variación con la actividad geomagnética

El Sol emite permanentemente, además de radiación electromagnética, partículas cargadas. Estas últimas constituyen lo que se conoce como viento solar, que emana de la atmósfera solar superior, es decir de la corona del Sol. La actividad geomagnética se refiere a la actividad de origen solar que afecta al campo magnético de la Tierra y que está relacionada con las partículas cargadas del viento solar (Matzka et al., 2021). Este viento se intensifica, con incrementos en su velocidad y flujo de partículas cargadas, y estas últimas penetran en la magnetosfera de la Tierra dando lugar a tormentas geomagnéticas e induciendo corrientes eléctricas que son las que finalmente dan lugar a las perturbaciones del campo magnético terrestre. Esta actividad, al igual que la solar, se la caracteriza con índices, denominados índices de actividad geomagnética.

La actividad geomagnética afecta a termosfera y a la ionosfera, y en particular la densidad electrónica, y por lo tanto a foF2. A diferencia del efecto de la actividad solar, donde un aumento del flujo EUV solar por aumento de la actividad solar produce, en general, un incremento de la ionización y de foF2 en casi todas las condiciones, y viceversa, la respuesta ionosférica ante perturbaciones geomagnéticas no es lineal ni uniforme.

Durante tormentas geomagnéticas, la ionosfera puede experimentar tanto aumentos (perturbaciones positivas) como disminuciones (perturbaciones negativas) de foF2, dependiendo de factores como la hora local, la latitud geográfica, la estación del año y la fase de la tormenta. En general, durante la fase principal de una tormenta severa, los valores de foF2 tienden a aumentar durante el día y a disminuir durante la noche, mientras que en la fase de recuperación puede observarse el comportamiento opuesto. Estas variaciones se explican principalmente por cambios en la composición y circulación de la termosfera, inducidos por la deposición de energía en las regiones de altas latitudes. Dicha energía, proveniente de corrientes magnetosféricas y de la precipitación de electrones, altera el balance entre producción, pérdida y transporte de ionización. En latitudes altas, las perturbaciones negativas son más frecuentes,

mientras que en las regiones ecuatoriales suelen observarse respuestas positivas. Las latitudes medias representan una zona de transición, donde la respuesta puede variar según la estación y la intensidad de la actividad geomagnética.

Por lo tanto, la relación entre la actividad geomagnética y foF2 es altamente variable y dependiente de las condiciones geofísicas locales, reflejando la compleja interacción entre la ionosfera, la termosfera y la magnetosfera. Sin embargo, en los modelos de regresión lineal de foF2 se suele incluir a los índices de actividad geomagnética de igual manera que los de actividad solar.

2.7 Índices de la actividad geomagnética

Indicadores clásicos de la actividad geomagnética, y que se derivan de mediciones del campo magnético que realizan observatorios en la superficie de la Tierra, son los índices Kp y Ap. Están además los índices aa y Dst, también muy usados pero que no describiremos ni se utilizarán en este trabajo. El registro de Kp y Ap comienza en el año 1932. Sus valores se pueden obtener libremente del Centro Mundial de Datos (World Data Center, WDC) para Geomagnetismo de Kyoto (<https://wdc.kugi.kyoto-u.ac.jp/index.html>).

Kp se obtiene como el promedio del índice K que corresponde a la perturbación magnética en las componentes horizontales del campo que se registra en 13 estaciones ubicadas en el mapa de la Figura 2.5. Se mide cada 3 horas (es decir que se registran 8 valores por día) y su escala es cuasi-logarítmica con un rango entre 0 y 9. El nombre viene de la palabra en alemán “planetarische Kennziffer” (índice planetario). Este índice mide la radiación de partículas del Sol a través del efecto que producen en el campo geomagnético. Se lo considera un proxy del input de energía del viento solar (Matzka et al., 2021).

Ap corresponde al promedio diario del índice ap. Este último corresponde a la linealización del índice Kp registrado cada 3 horas y se obtiene a partir de una tabla de equivalencias, como se muestra en la Tabla 2.1.

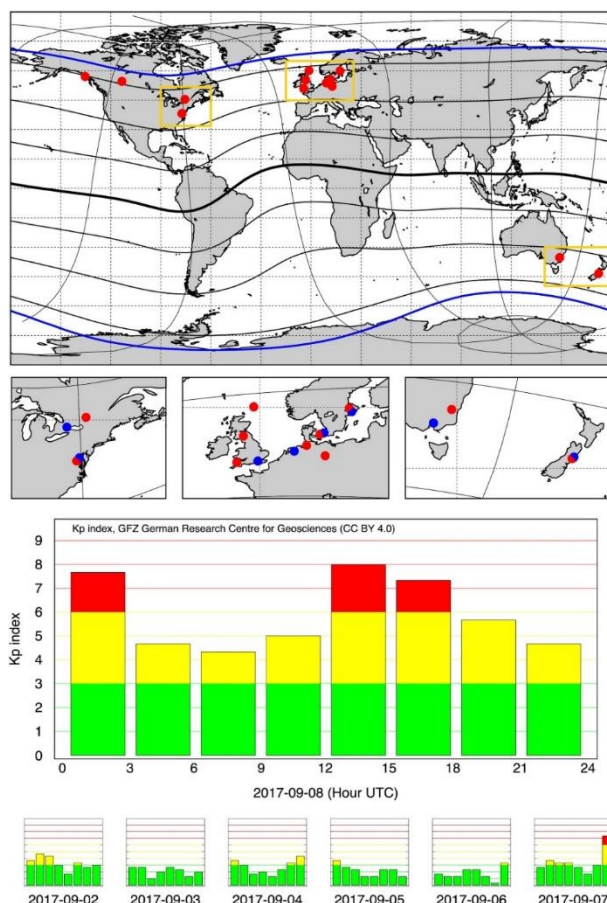


Figura 2.5 Paneles superiores: Mapa donde se indican las 13 estaciones actuales que se utilizan para estimar Kp (puntos rojos), con coordenadas geográficas (línea de puntos), coordenadas cuasi-dipolares (líneas llenas), el ecuador geomagnético (línea negra resaltada, 0° latitud geomagnética cuasi-dipolar), y zonas aurorales (líneas azules, +67° y -67° latitud geomagnética cuasi-dipolar). Los rectángulos amarillos indican las áreas agrandadas en los mapas regionales donde se marcan las estaciones que funcionan en la actualidad (rojo) y las previas que ya dejaron de funcionar (azul). Paneles inferiores: Gráficos que se proveen diariamente desde 1932, del índice Kp. Como ejemplo se muestra el 8 de septiembre de 2017, y 6 días previos. Fuente: Matzka et al. (2021).

Tabla 2.1 Equivalencia entre valores de Kp y ap. [Fuente: GFZ German Research Centre for Geosciences, <https://www.gfz-potsdam.de/en/kp-index>). ap tiene unidades de 2 nT.

Kp	0o	0+	1-	1o	1+	2-	2o	2+	3-	3o	3+	4-	4o	4+
ap	0	2	3	4	5	6	7	9	12	15	18	22	27	32
Kp	5-	5o	5+	6-	6o	6+	7-	7o	7+	8-	8o	8+	9-	9o
ap	39	48	56	67	80	94	111	132	154	179	207	236	300	400

Los índices de actividad geomagnética varían también con la actividad solar, pero con algunas diferencias. En la Figura 2.6 se grafica la serie de promedio mensual del índice Ap, que es el que se seleccionó en este trabajo para introducir en los modelos para interpolar los datos de foF2.

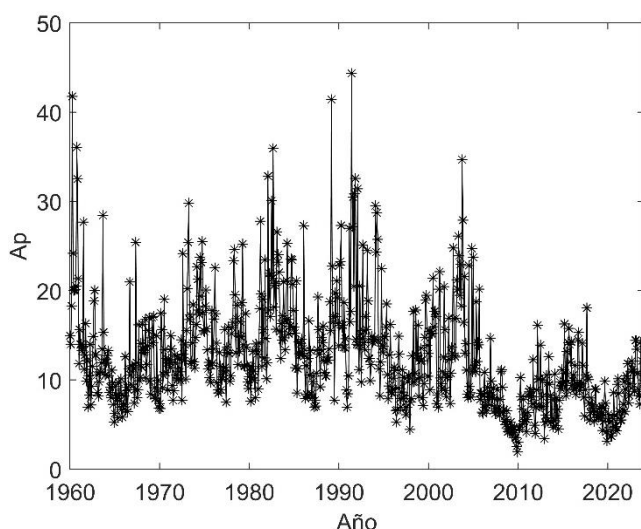


Figura 2.6 Series del promedio mensual del índice Ap de actividad geomagnética en el período enero 1960-diciembre 2023.

2.8 Índice global ionosférico IG

El índice global ionosférico, IG, propuesto por Liu et al. (1983), ha sido desarrollado en base a datos de ionosonda e incluye efectos que no son contemplados por los proxies solares, como por ejemplo el efecto de saturación de la ionosfera. La saturación ocurre en algunas estaciones ionosféricas cuando los índices de actividad solar exceden cierto valor, más allá del cual la ionosfera parece no responder al aumento de radiación EUV solar (Balan et al. 1994; Liu et al. 2006; Elias, 2014), por lo que se "rompe" la linealidad entre foF2 y el proxy solar considerado.

Este índice es distribuido ahora por el United Kingdom Solar System Data Centre (UKSSDC) (<https://www.ukssdc.ac.uk/>). En la Figura 2.7 se grafica la serie de promedio mensual de IG. Notar que toma valores negativos.

Originalmente, el índice se basaba en 11 estaciones de referencia, pero actualmente se determina a partir de cuatro estaciones: dos en el Hemisferio Sur (Port Stanley y Canberra) y dos en el Hemisferio Norte (Kokubunji y Chilton). Esta reducción ha limitado su capacidad para representar fielmente las condiciones ionosféricas globales. Sin embargo, para describir las variaciones del ciclo solar en la ionosfera, el IG mostrando en general un mejor desempeño que los índices SN y F10.7. Su principal ventaja radica en que, al derivarse de mediciones ionosféricas, incorpora efectos dinámicos que no están presentes en los índices puramente solares. Es además utilizado por el modelo IRI para predecir foF2.

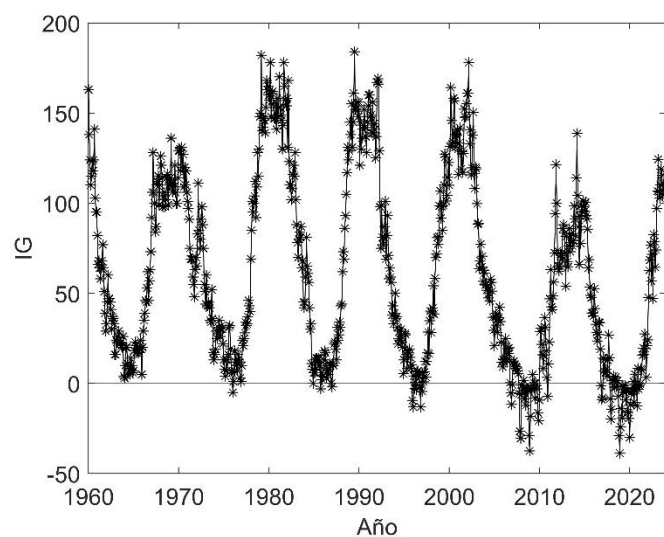


Figura 2.7 Series del promedio mensual del índice IG en el período enero 1960-diciembre 2023.

Capítulo 3. Métodos de interpolación basados en ML

Existen numerosos métodos de interpolación para completar datos faltantes. En este capítulo se describirán los métodos que se basan en técnicas de ML, cuyo desempeño se evaluará en esta Tesis para interpolar el parámetro ionosférico seleccionado.

3.1 ¿Qué es el "Machine Learning"?

El ML o aprendizaje automático es una subdisciplina de la inteligencia artificial cuyo objetivo es desarrollar algoritmos capaces de aprender a partir de los datos, sin estar explícitamente programados para realizar una tarea específica. En lugar de depender de reglas fijas definidas por un humano, los sistemas de ML construyen su propio conocimiento inductivo al identificar patrones y relaciones presentes en conjuntos de datos, para luego realizar predicciones o tomar decisiones sobre datos nuevos o no observados previamente (Mishra, 2020).

Los orígenes históricos del ML se remontan a 1959, cuando Arthur Samuel, un pionero en la computación, desarrolló en IBM un programa capaz de jugar al juego de damas. El sistema evaluaba cada posición del tablero asignándole una puntuación que representaba su probabilidad de llevar a una victoria. Estas puntuaciones no eran estáticas: el programa se entrenaba a sí mismo jugando contra versiones anteriores de sí mismo, y con cada iteración su desempeño mejoraba. Esta capacidad de mejorar a partir de la experiencia marcó el nacimiento formal del aprendizaje automático como campo de investigación.

Desde entonces, el ML ha evolucionado sustancialmente gracias a avances en algoritmos, en capacidad de cómputo, y en la disponibilidad masiva de datos provenientes de fuentes como redes sociales, sensores, imágenes satelitales, registros médicos o datos atmosféricos, entre otros.

En términos técnicos, un sistema de ML puede definirse como un conjunto de algoritmos basados en principios estadísticos y matemáticos, capaces de analizar grandes volúmenes de datos para extraer patrones subyacentes y generar modelos predictivos. Una característica fundamental de estos sistemas es su capacidad de generalización: es decir, su habilidad para aplicar lo aprendido en un conjunto de datos a casos nuevos no vistos durante el entrenamiento.

Es importante distinguir los sistemas de ML de los sistemas basados en reglas. En estos últimos, un humano define explícitamente las reglas de funcionamiento mediante estructuras del tipo if-then-else, es decir, instrucciones condicionales que se ejecutan en una secuencia determinada. Este enfoque, aunque útil en algunos contextos, puede no funcionar bien cuando se trata de datos complejos, ruidosos o de alto volumen. Por el contrario, un sistema de ML no requiere que un experto identifique y codifique manualmente los patrones: el propio sistema los descubre durante el proceso de entrenamiento y, además, puede seguir aprendiendo y ajustándose a medida que se expone a nuevos datos.

En la actualidad, el ML se ha convertido en una herramienta indispensable en múltiples disciplinas científicas y tecnológicas. Su aplicación va desde el reconocimiento de imágenes y voz, hasta la predicción de series temporales, la medicina personalizada, la conducción autónoma, y, particularmente en el caso de esta tesis, el análisis y la interpolación de datos ionosféricos. Esta capacidad de aprender de los datos y adaptarse a nuevas condiciones convierte al ML en una técnica especialmente poderosa frente a fenómenos complejos, no

lineales y de difícil modelado mediante métodos tradicionales (Camporeale et al., 2018; Camporeale, 2019; Mishra, 2020; Muzaffer & Senturk, 2022).

3.2 Regresión lineal

La regresión lineal es uno de los algoritmos fundamentales del ML y constituye la base del análisis predictivo. Su objetivo es modelar la relación entre una variable dependiente (resultado) y una o más variables independientes (predictores), permitiendo prever resultados a partir de datos de entrada. Esta técnica ayuda a identificar tendencias y relaciones, ofreciendo información clave para la toma de decisiones en diversos sectores.

En el contexto del aprendizaje supervisado, la regresión lineal aprende a partir de datos "etiquetados" (es decir, con entradas y salidas conocidas) para predecir resultados en datos nuevos. Se trata de un método de clara interpretación, ya que los coeficientes del modelo reflejan el impacto de cada predictor sobre el resultado, lo cual facilita la toma de decisiones basada en datos. Se puede definir entonces a la regresión lineal como un algoritmo de aprendizaje supervisado que predice valores modelando una relación lineal entre variables. El modelo intenta ajustar una línea recta que represente la mejor relación posible entre los predictores y la variable objetivo.

Es el modelo base, esencial para entender algoritmos más avanzados de ML como los que se verán más adelante. Hay dos tipos de modelos de regresión lineal: simple y múltiple. La regresión lineal simple considera una única variable independiente, mientras que la regresión lineal múltiple incorpora varias variables independientes para explicar el resultado.

Ambos tipos buscan ajustar una línea recta que describa la relación entre predictores y la variable objetivo, pero se diferencian en la complejidad y número de entradas que manejan.

En el caso de este trabajo, la variable objetivo o dependiente es foF2, y las variables independientes pueden ser el mes para incluir la estacionalidad, un proxy de actividad solar y/o un índice de actividad geomagnética.

La regresión lineal permite modelar la relación entre dos variables mediante una ecuación lineal de la forma:

$$y = \alpha_0 + \alpha_1 x \quad (3.1)$$

donde α_0 y α_1 son los coeficientes de regresión, también conocidos como parámetros del modelo, que se estiman generalmente mediante el método de mínimos cuadrados. En esta expresión, x representa la variable independiente y " y " la variable dependiente. En el contexto del presente trabajo, considerando por ejemplo las series para cada hora y mes por separado, a lo largo del período 1960-2023, estas variables corresponderían, respectivamente, a un proxy de EUV y foF2. No hace falta tener en cuenta en el modelo la variación diurna ni la estacional porque al separar en horas y meses estas periodicidades desaparecen y solo quedan los ciclos asociados a la actividad solar y/o geomagnética, más el ruido natural de las series de tiempo.

La simplicidad de este método y su bajo costo computacional permiten establecer una línea base que sirve como referencia para evaluar el rendimiento de algoritmos más sofisticados en la tarea de estimar valores faltantes en series temporales de datos ionosféricos

La regresión múltiple es una generalización del modelo de regresión lineal simple, que permite incorporar múltiples variables independientes para explicar la variabilidad de un parámetro

dependiente. Este enfoque resulta especialmente útil cuando el fenómeno a estudiar está influenciado por más de un forzante físico, lo cual es común en sistemas complejos como la ionosfera. En el caso particular de foF2 además del proxy de radiación solar en el rango EUV utilizado en la regresión lineal simple, se pueden considerar otros factores relevantes, como un índice de actividad geomagnética. O, por ejemplo, si se detecta una falta de linealidad con algún forzante, se pueden agregar términos donde se le aplica alguna función al forzante x , como x^2 . La expresión general para un modelo de regresión múltiple es:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n \quad (3.2)$$

donde x_n representan las variables independientes (tales como proxies solares, índices geomagnéticos o datos de estaciones ionosféricas vecinas, proxies elevados a alguna potencia distinta de 1), y α_n son los coeficientes de regresión que cuantifican la contribución de cada una de estas variables a la variable dependiente y , que en este caso es foF2. Este tipo de modelización resulta útil para capturar la influencia combinada de diferentes forzantes sobre la ionosfera, permitiendo una mejor estimación de los valores faltantes o la predicción de comportamientos futuros.

3.3 Regresión por Bosques Aleatorios (Random Forest Regression)

La Regresión por Bosques Aleatorios (Random Forest Regression, RFR) es una técnica de ML basada en el concepto de aprendizaje por conjunto (ensemble learning), propuesta por Leo Breiman en 2001 (Breiman, 2001). Este enfoque consiste en la construcción de múltiples modelos individuales que en este caso son árboles de decisión, cuya combinación permite obtener una predicción más robusta, precisa y generalizable que la que podría lograrse con un único modelo. La Figura 3.1 muestra de manera esquemática un árbol de decisión.

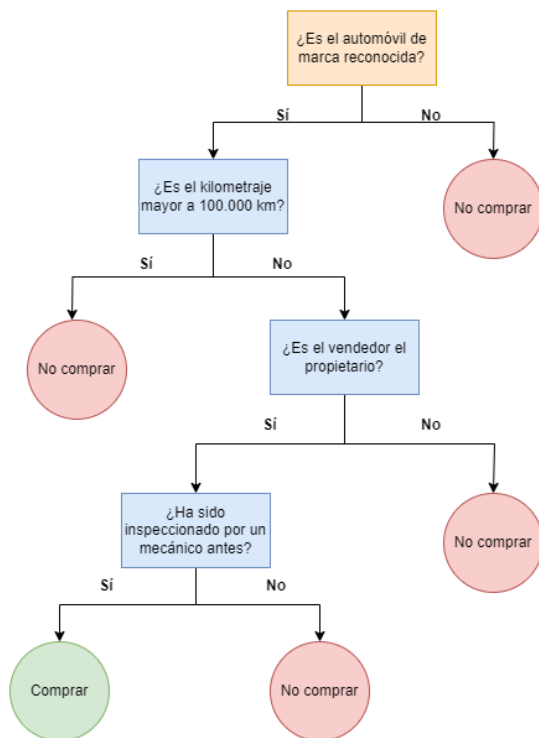


Figura 3.1 Esquema de un árbol de decisión: el bloque amarillo es el nodo raíz, los bloques azules son los nodos de decisión, los círculos verdes y rojos representan las hojas, en las que se decide la decisión final según cada respuesta a cada nodo. Fuente: <https://www.educaopen.com/digital-lab/blog/inteligencia-artificial/arboles-de-decision>

Durante el proceso de entrenamiento, se construyen T árboles de decisión independientes utilizando subconjuntos aleatorios del conjunto de datos de entrenamiento, tanto en las observaciones como en las variables (técnica conocida como bootstrap aggregating o bagging). Cada uno de estos árboles realiza una predicción individual, y la predicción final del modelo se obtiene promediando las salidas de todos los árboles:

$$\hat{y} = \frac{1}{T} \sum f_t(x) \quad (3.3)$$

Donde:

- $\hat{y}$ es la predicción final del modelo de RFR para una entrada x
- T es el número total de árboles en el bosque
- $f_t(x)$ representa la predicción del árbol número t

Cada árbol de decisión individual funciona dividiendo recursivamente el espacio de características en regiones homogéneas respecto al valor de la variable dependiente. Estas divisiones se realizan seleccionando, en cada nodo del árbol, la variable y el umbral que minimicen la suma del error cuadrático medio (MSE, por sus siglas en inglés) dentro de los subconjuntos generados. Así, el árbol se estructura jerárquicamente, donde los nodos internos representan decisiones sobre variables independientes, y las hojas contienen los valores predichos.

Este enfoque presenta varias ventajas importantes:

- Reduce el riesgo de sobreajuste, común en los árboles de decisión individuales.
- Es relativamente robusto frente a datos ruidosos o valores atípicos.
- Permite obtener medidas de importancia de variables, ya que los árboles priorizan aquellas características que más contribuyen a la reducción del error.

La RFR ha demostrado ser una técnica eficaz para problemas con relaciones no lineales complejas, interacciones entre variables, o conjuntos de datos con una gran cantidad de atributos. En el contexto de este trabajo, el modelo se emplea como uno de los métodos principales para la estimación de valores faltantes en la serie temporal de la frecuencia crítica de la capa F2 (foF2), aprovechando su capacidad para modelar relaciones no lineales y capturar patrones complejos en los datos ionosféricos.

3.4 Regresión por Gradient Boosting (Gradient Boosting Regression)

La Regresión por Gradient Boosting (GBR) es una técnica de aprendizaje supervisado desarrollada por Jerome Friedman en 2001 (Friedman, 2001). Esta metodología se basa en el principio de boosting, que consiste en construir un modelo fuerte a partir de una secuencia de modelos débiles (comúnmente árboles de decisión de baja profundidad), los cuales se agregan de manera secuencial para corregir los errores cometidos por el conjunto acumulado hasta el momento. En términos funcionales, el modelo de GBR busca aproximar la relación subyacente entre las variables independientes y dependientes como una suma ponderada de modelos base:

$$\hat{y} = \sum \gamma_m h_m(x) \quad (3.4)$$

Donde:

- $\hat{y}$ representa la predicción del modelo
- M es el número total de iteraciones o etapas
- $h_m(x)$ es el modelo base (habitualmente un árbol de decisión) en la etapa m
- γ_m es la tasa de aprendizaje o learning rate, que modula la contribución de cada modelo al conjunto.

El procedimiento comienza con una predicción inicial, que típicamente es la media de los valores de la variable dependiente. Luego, en cada iteración m , se entrena un nuevo árbol para predecir el residual o el gradiente del error de la función de pérdida, es decir, la diferencia entre los valores reales y los valores predichos por el modelo acumulado hasta esa instancia. De este modo, cada árbol posterior está enfocado en los casos donde el modelo previo tuvo mayor error, lo que permite un ajuste fino progresivo del modelo total. Una característica fundamental del GBR es su enfoque secuencial: cada nuevo modelo se construye en función de los errores del anterior. Esto lo diferencia conceptualmente de la RFR, en la que todos los árboles son entrenados de manera independiente y sus salidas se promedian. La Figura 3.2 muestra un esquema de comparación entre GBR y RFR.

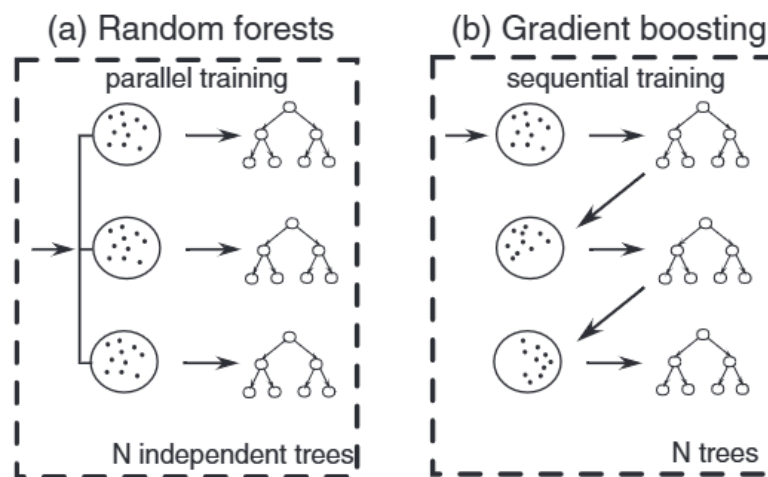


Figura 3.2 Comparación entre (a) Random Forest (RFR) y (b) Gradient Boosting (GBR). En el método de Random Forest, se construyen N modelos independientes (árboles) en paralelo a partir de subconjuntos aleatorios del conjunto de datos de entrada. En cambio, en el método de Gradient Boosting, cada nuevo árbol se construye a partir de los residuos del ensamble anterior. Fuente: Kowalek et al. (2019)

Mientras que RFR se beneficia de la paralelización computacional (los árboles pueden entrenarse simultáneamente), GBR no permite esta ventaja, ya que el entrenamiento de cada nuevo árbol depende del error residual del conjunto acumulado hasta ese punto. A pesar de su mayor costo computacional, el método de GBR suele ofrecer mayor precisión predictiva, especialmente en conjuntos de datos con patrones complejos y relaciones no lineales. El parámetro de tasa de aprendizaje γ permite controlar el grado de ajuste del modelo: valores pequeños de γ hacen que el modelo aprenda más lentamente, pero con menor riesgo de sobreajuste. En el presente trabajo, la GBR se emplea como uno de los métodos de interpolación

para la estimación de valores faltantes en la serie temporal de la frecuencia crítica de la capa F2 (foF2). Su capacidad para capturar estructuras complejas y refinar progresivamente la predicción resulta especialmente útil en series con variabilidad no lineal o influencias múltiples como las que caracterizan a los fenómenos ionosféricos.

3.5 Redes Neuronales (Neural Networks)

Las Redes Neuronales (Neural Networks) son también una técnica de ML que extiende los modelos de regresión mediante la combinación de numerosas unidades interconectadas organizadas en capas sucesivas. Esta arquitectura jerárquica permite capturar relaciones no lineales complejas entre las variables de entrada y salida, superando las capacidades descriptivas de los modelos de regresión tradicionales. Cuando estas arquitecturas incluyen múltiples capas ocultas, como se muestra esquemáticamente en la Figura 3.3, se las engloba dentro del campo conocido como Aprendizaje Profundo (Deep Learning).

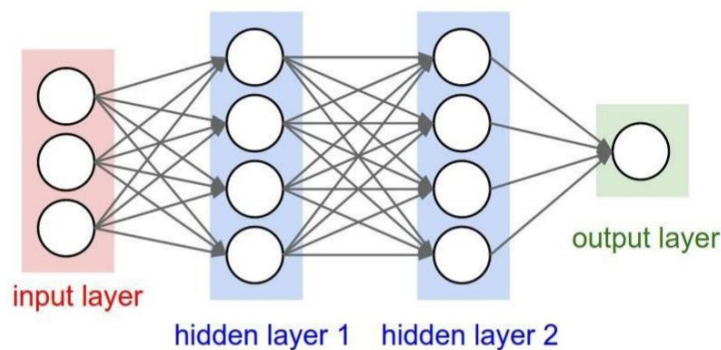


Figura 3.3 Estructura de red neuronal con una primera capa de entrada (input layer), dos capas ocultas (hidden layer) y una capa de salida con una sola neurona (output layer).

En este contexto, una neurona es simplemente una función, es decir que se trata de "algo" que recibe uno o más valores de entrada, le aplica una determinada lógica y genera un resultado. Las neuronas, a su vez, están interconectadas entre sí. Cada neurona recibe entradas de otras neuronas o de fuentes externas, y producen su salida que, a su vez, es enviada a otras neuronas o a la salida final de la red. En una red neuronal artificial, las neuronas se organizan en capas, y la información fluye de la capa de entrada a través de una o más capas intermedias (conocidas como capas ocultas) hasta la capa de salida (ver Figura 3.3). Cada capa está compuesta por varias neuronas, y la interconexión entre las neuronas permite que la red aprenda a reconocer patrones y realizar tareas específicas, como clasificar imágenes o predecir valores numéricos, como sería en el caso de este trabajo de Tesis.

En este trabajo se utilizó una red neuronal artificial de tipo MLP o red densamente conectada de propagación hacia adelante (feed forward) (ver esquema de la Figura 3.4b). Estructuralmente, este tipo de arquitectura se fundamenta en la interconexión secuencial de capas compuestas por unidades de procesamiento elemental, es decir "neuronas", donde la información fluye en un único sentido, desde la capa de entrada hasta la capa de salida. Cada neurona en una capa determinada recibe como variables de entrada las activaciones ponderadas

de la capa previa, les aplica un sesgo y, posteriormente, transforma dicha suma lineal mediante una función de activación. Este mecanismo confiere al modelo la capacidad de actuar como un aproximador universal de funciones, permitiéndole mapear y aprender relaciones complejas, multidimensionales y altamente no lineales entre variables físicas predictoras y una variable de interés, sin necesidad de asumir a priori un modelo analítico preestablecido.

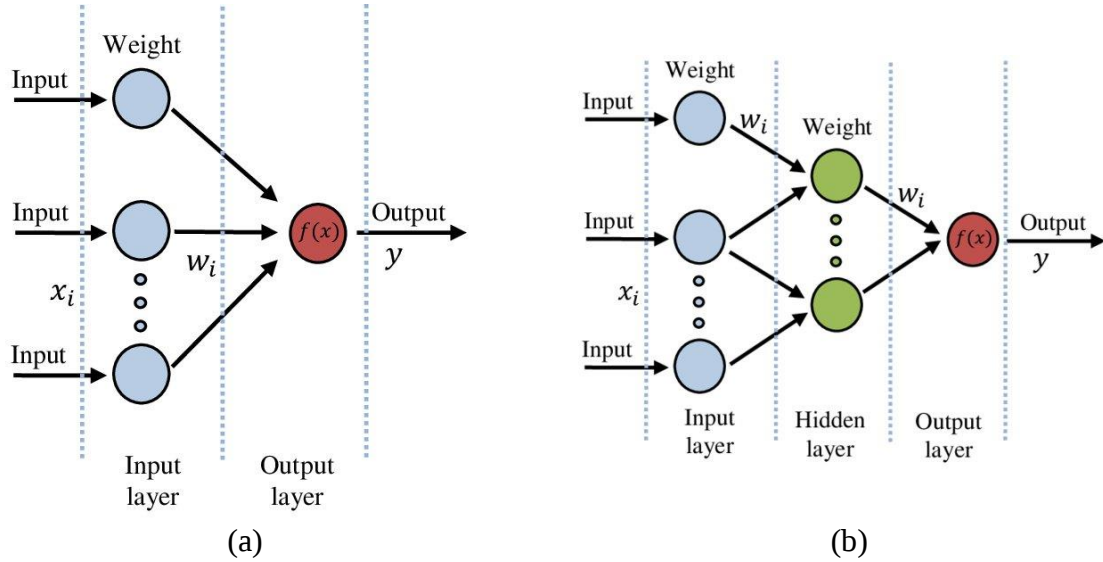


Figura 3.4 (a) Arquitectura de un perceptrón de una capa. (b) Arquitectura de un perceptrón multicapa, con una capa oculta.

La red diseñada en este trabajo se configuró como un regresor profundo estructurado en cinco capas diferenciadas: una capa de entrada, tres capas ocultas de procesamiento intermedio y una capa de salida. La dimensionalidad de la entrada está estrictamente determinada por el número de características físicas y temporales seleccionadas a partir de los registros de proxies solares e índices ionosféricos. La información normalizada ingresa al núcleo de la red a través de la primera capa oculta, la cual consta de 64 neuronas. A esta le suceden de forma consecutiva dos capas ocultas adicionales con un esquema de reducción progresiva de dimensionalidad, compuestas por 32 y 16 neuronas, respectivamente. Esta reducción geométrica en el número de nodos actúa como un filtro abstracto de características, forzando a la red a condensar la información y extraer los patrones físicos más robustos a medida que se profundiza en el modelo. Finalmente, el flujo converge en una capa de salida unitaria, conformada por una única neurona dedicada a proyectar la predicción correspondiente al parámetro ionosférico crítico foF2.

La dinámica no lineal del modelo y su proceso de optimización están gobernados por funciones matemáticas específicas en cada nivel de la arquitectura. Para las tres capas ocultas, se implementó la función de activación de Unidad Lineal Rectificada (ReLU, por sus siglas en inglés), definida matemáticamente como $f(x) = \max(0, x)$. La elección de ReLU resulta fundamental para mitigar de manera eficiente el problema del desvanecimiento del gradiente (vanishing gradient) durante la retropropagación en redes profundas, acelerando sustancialmente la convergencia del entrenamiento al mantener derivadas constantes para valores de excitación positivos. Por el contrario, la neurona de la capa de salida utiliza una

función de activación lineal, una condición necesaria en problemas de regresión que permite al modelo proyectar valores continuos en todo el espectro de los números reales, sin saturación ni acotamiento artificial de la escala de la variable física a predecir.

El aprendizaje y ajuste de los pesos sinápticos y sesgos de la red se lleva a cabo mediante la minimización de una función de costo basada en el Error Cuadrático Medio (MSE) e inicio He.

Capítulo 4. Datos y Metodología

En este capítulo se describirán las bases de datos y fuentes de foF2, junto con la fuente particular que se utilizó para obtener la serie analizada en esta Tesis que corresponde a la estación de Juliusruh, en Alemania, y que es una de las más completas y homogéneas, en cuanto a calidad, que existe. También se describirán la metodología empleada en la generación de huecos para analizar el desempeño de los métodos de ML para interpolación, y los estadísticos de comparación usados para evaluar el desempeño de cada uno de estos métodos.

4.1 Fuentes de datos de foF2

Los parámetros o características ionosféricas que se extraen de los ionogramas, entre los que se encuentra foF2, están disponibles libremente desde hace muchos años en los Centros de Datos Mundiales de Física Solar Terrestre (World Data Centers -WDC- for Solar-Terrestrial Physics -STP-), donde son almacenados en bases de datos con un formato estandarizado. Todos ellos disponen de datos de ionosonda de diferentes estaciones. Los valores de las ionosondas se registraban por hora, y los de digisonda se registran cada 5, 10 o 15 minutos. Algunos WDC son:

- WDC operado por el "Australian Space Weather Forecasting Centre" (ASWFC), miembro del International Science Council (ISC) World Data System, ubicado en Sydney, Australia (disponible en https://www.sws.bom.gov.au/World_Data_Centre)
- WDC del United Kingdom Solar System Data Centre (UKSSDC) (disponible en https://www.ukssdc.ac.uk/wdcc1/data_menu.html)
- WDC de Moscú, Rusia (disponible en <http://www.wdcb.ru/stp/index.en.html>)
- WDC de Japón, NICT (National Institute of Information and Communications Technology) (disponible en <https://wdc.nict.go.jp/Ionosphere/en/index.html>)

Desde 2002 existe el Global Ionosphere Radio Observatory (GIRO) (Reinisch & Galkin, 2011) que reúne datos de ionosondas Digisondas instaladas en más de 80 estaciones alrededor del mundo (ver mapa de Figura 4.1). A través de su portal (<https://giro.uml.edu/index.html>), brinda acceso público a millones de registros de ionogramas y otros parámetros de utilidad para telecomunicaciones y alertas de Space Weather entre otros, muchos de ellos en tiempo real. Este recurso constituye una base fundamental para estudios de la respuesta ionosférica a eventos de Space Weather y para la mejora de modelos ionosféricos como el IRI.

Los registros de parámetros ionosféricos vienen acompañados de letras calificativas y descriptivas definidas en el "URSI Handbook of Ionogram Interpretation and Reduction" (Piggott & Rawer, 1978) (URSI corresponde a las siglas de Union Radio-Scientifique Internationale, en Inglés: International Union of Radio Science). Estas son:

Letras calificativas:

A y E – Menor que

D – Mayor que

I – Valor faltante reemplazado por un valor interpolado

J – Característica de la componente ordinaria deducida de la extraordinaria

M – Interpretación dudosa porque no es posible distinguir entre las componentes ordinaria y extraordinaria

O – Característica de la componente extraordinaria deducida de la ordinaria

T – Valor determinado a partir de una secuencia de observaciones, siendo la observación real inconsistente o dudosa

U – Valor numérico incierto o dudoso

Z – Medición deducida de la tercera componente magneto-electrónica

Letras descriptivas:

A – Medición influida por, o imposible debido a, la presencia de una capa inferior delgada (por ejemplo, E esporádica).

B – Medición influida por, o imposible debido a, absorción en las cercanías de la frecuencia mínima

C – Medición influida por, o imposible debido a, alguna causa no ionosférica.

D – Medición influida por, o imposible debido al límite superior del rango de frecuencias normal

E – Medición influida por, o imposible debido al límite inferior del rango de frecuencias normal

F – Medición influida por, o imposible debido a la presencia de ecos difusos (spread echoes)

G – Medición influida por, o imposible debido a que la densidad de ionización de la capa reflectora es demasiado baja para obtener una medida precisa

H – Medición influida por, o imposible debido a la presencia de estratificación

K – Presencia de una capa E generada por precipitación de partículas, y con foE mayor que lo normal

L – Medición influida por, o imposible porque la traza no presenta un punto de inflexión suficientemente definido entre las capas

M – Interpretación dudosa porque no es posible distinguir entre las componentes ordinaria y extraordinaria

N – Las condiciones son tales que la medición no puede interpretarse

O – La medición se refiere a la componente ordinaria

P – Medición influida por, o imposible debido a la presencia de anomalías polares.

Q – Medición influida por, o imposible debido a la presencia de dispersión en rango (range spread)

R – Medición influida por, o imposible debido a atenuación en las cercanías de la frecuencia crítica

S – Medición influida por, o imposible debido a interferencia o ruidos atmosféricos

T – Valor determinado a partir de una secuencia de observaciones, siendo el valor u observación real inconsistente o dudoso

V – Traza bifurcada que puede influir en la medición

W – Medición influida por, o imposible debido a que el eco está fuera del rango de alturas registrado

X – La medición se refiere a la componente extraordinaria

Y – Fenómeno de "Lacuna": ausencia de ecos en determinadas frecuencias del ionograma

Z – Presencia de una tercera componente magneto-electrónica



Figura 4.1 Mapa de estaciones que integran GIRO, operativas y proyectadas. Estos sitios proveen además datos en tiempo real que son asimilados en modelos ionosféricos. Fuente: <https://giro.uml.edu/index.html>

El proyecto desarrollado por Damboldt y Suessmann tuvo como objetivo principal recopilar, verificar y estandarizar a nivel global las mediciones de las características ionosféricas foF2 y M(3000)F2. La finalidad de esta consolidación fue estructurar una base de datos unificada y consistente a largo plazo que permitiera validar y contrastar de manera empírica las predicciones de los modelos ionosféricos teóricos, como el modelo estático de la Unión Internacional de Telecomunicaciones (ITU), frente a observaciones históricas reales.

La recopilación de la información abarcó registros provenientes de los Centros Mundiales de Datos (WDC) de Japón, Reino Unido, Moscú, Boulder (EE. UU.) y Sídney (Australia), junto con la plataforma SPIDR (Space Physics Interactive Data Resource) y datos provistos por estaciones ionosonda individuales. Como punto de partida metodológico, se tomaron y expandieron los conjuntos de datos previamente depurados por Jürgen Bremer en sus estudios sobre tendencias ionosféricas.

Durante la fase de diagnóstico, se identificó una marcada incompatibilidad en los formatos de los archivos originales de los diferentes centros de datos. Entre las principales inconsistencias se destacaron marcas de tiempo erróneas o ambiguas (coexistencia de tiempo local y tiempo universal), discrepancias en las escalas de las unidades de medida (factores de multiplicación

inconsistentes en MHz), registros faltantes y duplicaciones con valores contradictorios para una misma estación y período.

Para resolver estas deficiencias, se implementaron rutinas de software diseñadas para homogeneizar toda la información temporal bajo el estándar de Tiempo Universal (UT). En lo respectivo al tratamiento estadístico, se estableció que el cálculo de las medianas mensuales requería de un umbral mínimo de observaciones válidas de al menos 10 días por mes para cada hora específica. Asimismo, se aplicaron filtros de control de calidad rigurosos para la detección y remoción de valores atípicos y para la resolución de discrepancias mediante la interpolación y el arbitraje de fuentes.

El resultado final del procesamiento consistió en una base de datos integrada y depurada, estructurada en archivos individuales por estación mediante su código URSI. Mediante la combinación sistemática de los repositorios disponibles, el documento final incrementó significativamente la densidad y continuidad de la información de las medianas mensuales, superando en cobertura y fiabilidad a cualquier base de datos ionosférica individual consultada previamente.

4.2 Estación seleccionada: Juliusruh

Para este trabajo de Tesis se seleccionó el parámetro foF2 medido en la estación de Juliusruh, y en particular se utilizó la serie de tiempo de los valores de la mediana mensual de cada hora del día, en el período 1960-2023 (64 años).

Se seleccionó esta estación ionosférica de Juliusruh (54.6°N, 13.4°E) por ser una de las estaciones con la base de datos más extensa con menos huecos, y con un buen control de calidad. Esto es indispensable ya que, para períodos tan extensos, los datos deben ser consistentes y homogéneos, lo que necesita de controles de calidad que la serie de Juliusruh posee. En un período de estas longitudes temporales es necesario que estén bien documentados los cambios de instrumentos y que haya una buena calibración con el equipo que se reemplaza para evitar "saltos artificiales" en la serie de tiempo. Debe tenerse en cuenta también el envejecimiento de los equipos y su efecto sobre el registro de datos (Lastovicka & Jelinek, 2019).

Se utilizó la serie de medianas mensuales dado que gran parte de los investigadores del LIANM analizan tendencias de largo plazo asociadas al efecto invernadero y a la variación secular del campo geomagnético, y utilizan este tipo de valores de foF2, que son provistos también por los WDC.

Antiguamente las medidas extraídas de los ionogramas se reportaban en tablas mensuales, como la que se muestra en la Figura 4.2, donde se reportan los valores horarios de cada día acompañados de dos letras: la calificativa y la descriptiva. Al final de la tabla se incluían 5 filas más: el número de valores diarios registrados (C), el valor de la mediana mensual (M) y los cuartiles superior e inferior (UQ y LQ), junto con el rango inter-cuartil (QR=UQ-LQ). Los valores de las medianas mensuales son de gran interés y uso porque la climatología de la ionosfera se analiza y estima en función de estos valores.

STATION ROME										LAT 41.8 N LONG 12.5 E										AUT. SWEEP 1 TO 20 MHz									
UNIVERSAL TIME										CHARACTERISTIC 00 FOF2										UNIT 0.1 MHz JAN 08									
000 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300																													
1	*	37	37	36	34	31	27	33	46	43	62	56	52	43	48	50	48	34	26	28	28	32	32	34	33				
2	*	30	31	31	30	23	21	J	27RJ	40R	50	48	66	45	52	45	58	49	J	31R	32	30	34	36	31	33			
3	*	32	33	34	31	29	27	30	45	42	49	58	56	42	40	55	44	36	22	27	29	33	26	I	28C	29			
4	*	31	29	32	32	31	31	32	J	48R	46	42	53	57	50	47	48	57	39	28	31	38	35	J	28R	33	34		
5	*	31Z	32Z	36	36	31	31	36	48	45	53	53	57	51	63	68	40	32	38	48	47	43	48	52	24	*			
6	*	48	52	55	50F	50	50	40	47	66	62	68	J	60R	58	52	66	52	51	33	33	37	32	32	35	33			
7	*	31	34	32	32	35	23	26	45	58	65	66	62	54	55	63	44	41	32	28	27	26	30	32	32				
8	*	31F	37	35	42	31	32	35	54	46	63	70	65	48	53	55	51	46	32	30	32	25	29	31	29				
9	*	26	28	23	25	24	22	28	J	49R	49	62	54	58	51	54	56	49	39	27	37	27	23	26	29	26			
10	*	27	29	34	36	42	J	32R	28	39	43	54	61	J	50R	61	55	59	49	45	36	J	27R	30	22	25	29	28	
11	*	29	28	27	24	24	22	26	44	45	55	51	56	52	I	52R	52	48	38	37	35	33	35	43	40	38V			
12	*	41	43	41	41	44	32	31	41	45	49	55	56	45	52	66	46	33	23	28	30	30	31	30	28				
13	*	28	29	27	25	20	23	30	39	44	54	58	63	51	57	47	43	47	43	38	33	30	35	36	35				
14	*	36	37	34	30F	34	31	29	41	58	53	J	60R	51	57	52	60	46	40	29	J	32A	37	25	28	30	30		
15	*	30	27	27	25	26	21	26	43	46	50	48	50	62	58	58	50	46	34	36	30	27	29	32	32				
16	*	33	32	32	31	32	25	21	42	43	52	70	54	50	53	54	45	40	32	J	28R	27	22	23	26	25			
17	*	27	26	J	29R	29	J	27R	23	J	28R	37	45	42	62	56	50	42	39	25	24	26	25	25	28	30			
18	*	30	33	32F	43	30	23	25	45	44	46	43	60	48	54	47	53	45	28	30	33	27	25	26Z	28				
19	*	28	28Z	31	27	26	26	23	I	35C	47	48	51	67	59	47	54	49	38	23	30	28	25	25	28	28			
20	*	28	28	29	28	J	26R	B	22	41	44	40	52	57	50	46	49	54	34	26	37	36	26	32	33	35			
21	*	35	35	35	33	29	23	26	41	45	48	54	51	51	47	61	42	36	32	31	J	29A	16	25	26	24			
22	*	26	29	25V	26	29	24	24	44	46	48	47	48	47	53	51	42	39	J	31A	28	25	28	35	30	24			
23	*	26	27	25	24	22	24	42	47	55	46	43	46	51	53	42	33	35	30	25	27	33	32	32	29				
24	*	31	29	26	J	26R	25	19	22	43	47	50	48	47	48	49	51	41	43	30	32	32	35	36V	35	35			
25	*	31	29	30V	32	J	27R	21	24	44	52	64	53	53	45	55V	53	43	36	J	32A	41	40	38	34	37F	31		
26	*	34	36	35	34	30	22	25	42	53	50	49	49	48	58	49	53	42	32	A	A	31	35	35F	32F				
27	*	29V	31	30	30	33	33	33	41	42	46	54	50	47	47	56	49	42	29	J	33R	32	30	31	33	30			
28	*	32	31	26	25	24	23	25	44	48	57	52	53	53	54	56	48	J	40R	27	26	27	29	28	30	J	28R		
29	*	28	26	25	J	25R	24	24	30	43	42	43	52	57	51	49	43	51	44	31	28	37	42	37	34	37			
30	*	37	41	41	41	41	37	34	39	J	47R	53	J	50R	47	52	55	50	48	36	23	27	33	30	27	28	27		
31	*	27	28	28	26	24	21	J	28R	45	42	45	37	51	52	47	48	53	39	27	26	30	32	29	30	29			
C	*	31	31	31	31	31	30	31	31	31	31	31	31	31	31	31	31	31	31	31	30	30	31	31	31	31			
M	*	31	31	31	31	30	29	23	28	43	46	50	53	54	51	52	54	48	39	31	30	31	30	30	32	30			
UQ	*	33	35	35	34	32	31	31	45	48	55	60	57	52	55	58	51	43	32	33	34	33	34	34	34	33			
LQ	*	28	28	27	26	25	22	25	41	44	48	50	50	48	47	50	43	36	27	28	28	25	26	29	28				
LR	*	5	7	8	8	7	9	6	6	4	4	7	10	7	4	8	8	7	5	5	6	8	8	5	5				

Figura 4.2 Página de los boletines ionosféricos mensuales que publicaban las estaciones de acuerdo a las reglas de URSI.

Los WDC proveen entonces, además de los datos horarios medidos, el registro de las medianas mensuales de los parámetros ionosféricos.

El uso de valores medianos mensuales de foF2 en lugar de promedios es un método recomendado para reducir el efecto de la actividad geomagnética en los datos ionosféricos, dado que la mediana es menos sensible a valores extremos que la media (Piggott & Rawer, 1972).

Se busca en general que el valor promedio horario de cualquier parámetro dado para cada mes sea representativo, es decir que sea un valor típico y cercano al valor más probable. Esto es difícil en la práctica, ya que la distribución de valores es casi siempre sesgada. Usualmente sucede que un número variable de valores se encuentran muy alejados del valor promedio más probable debido en general a la ocurrencia de tormentas geomagnéticas u otro tipo de perturbaciones a las cuales es sensible la ionosfera. En estos casos, la mediana resulta más representativa de las condiciones promedio que la media. Por lo tanto, se la ha adoptado como la medida estándar del valor promedio a utilizar internacionalmente en los parámetros ionosféricos que se registran a partir de los ionogramas. Y, por las mismas razones, para medir la variabilidad de los datos, en vez de usar la desviación standard, se utiliza el rango intercuartiles para mostrar la variabilidad de los datos en cada hora.

La representatividad de un valor promedio depende de la forma de la distribución de los datos originales individuales a partir de los cuales se estimó ese valor. Con las distribuciones sesgadas que suelen presentar los parámetros ionosféricos, la confiabilidad del promedio disminuye rápidamente cuando el número de observaciones es pequeño. Por lo tanto, es importante, siempre que sea posible, tratar de tener todos los valores del mes para computar la mediana o el cuartil. Esto no siempre ocurre, pero igual se calculan los valores de las medianas, y por eso se registra junto a este valor, el número de datos diarios considerados para su estimación.

Los datos de foF2 mediana mensual están disponibles en el WDC operado por Australia, en el período 1957-2023. La serie de medianas mensuales que se obtiene tiene datos faltantes en el período 1983-1990. Para obtener la base completa en el período considerado, se combinó con la base de datos de Damboldt & Suessmann (2012) (provista por la misma WDC) que da la serie completa de medianas mensuales de foF2 hasta 2009.

La Figura 4.3 muestra los valores de foF2 en todo el período considerado (1960-2023).

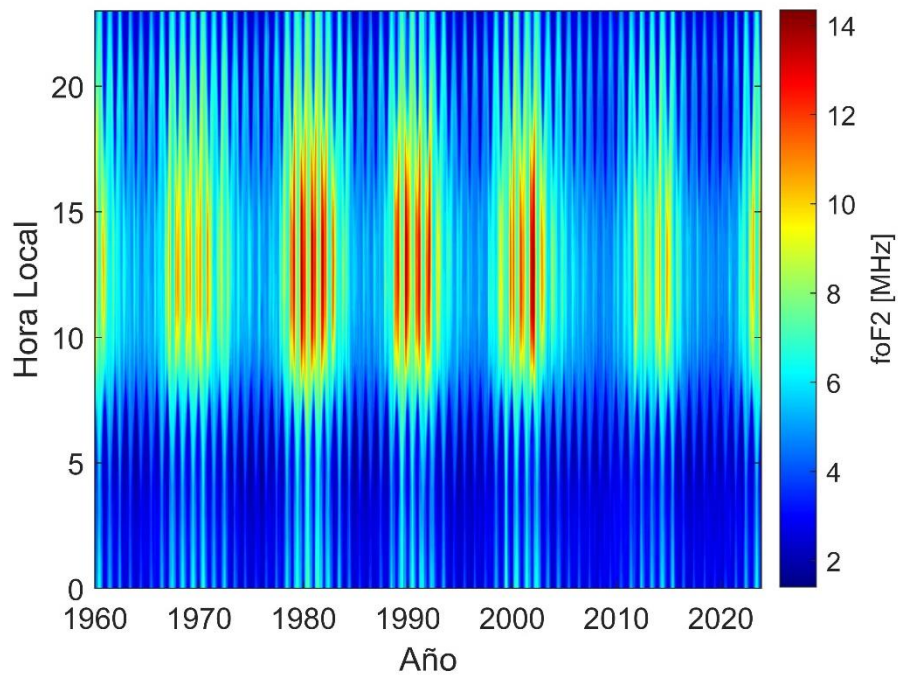


Figura 4.3 Mediana mensual de foF2 [MHz] medido en Juliusruh en el período 1960-2023.

La Figura 4.4 muestra una sección de la Figura 4.3 correspondiente a la hora 12 LT donde se ve claramente la variación estacional y la asociada a la actividad solar.

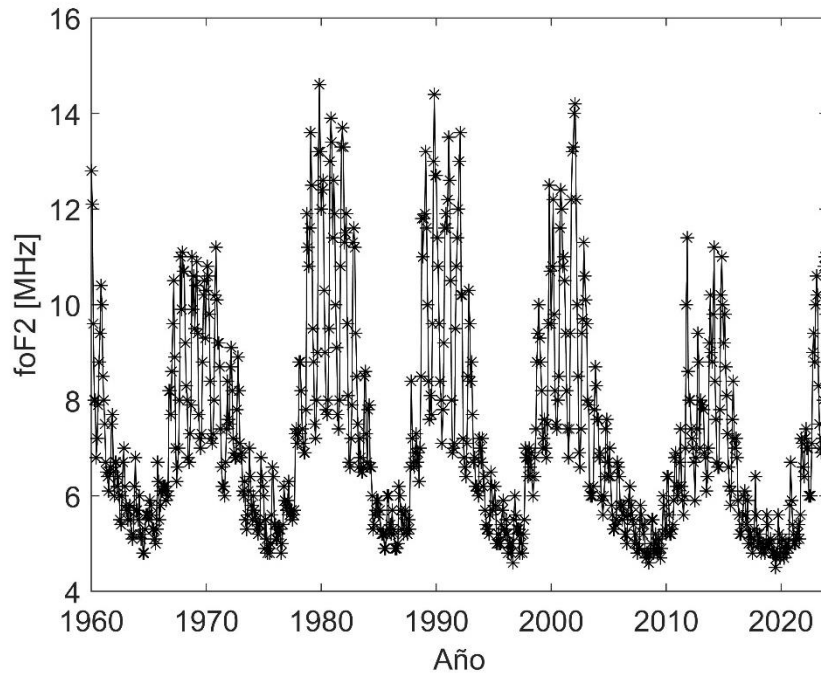


Figura 4.4 Mediana mensual de foF2 [MHz], para la hora 12 LT, medido en Juliusruh en el período 1960-2023.

En la Figura 4.5 se grafica el valor promedio de foF2 considerando todos los años para cada mes y hora. Se observa la variación estacional para cada hora, y claramente al mediodía local se nota que el valor máximo de foF2 no ocurre en los meses de verano sino en los equinoccios y meses de invierno. En la Figura 4.6 se grafica el valor promedio para cada hora considerando el período completo. Se observa en este caso, como es esperable por la hora donde se produce el máximo de radiación solar, el valor máximo de foF2 a las 12 LT.

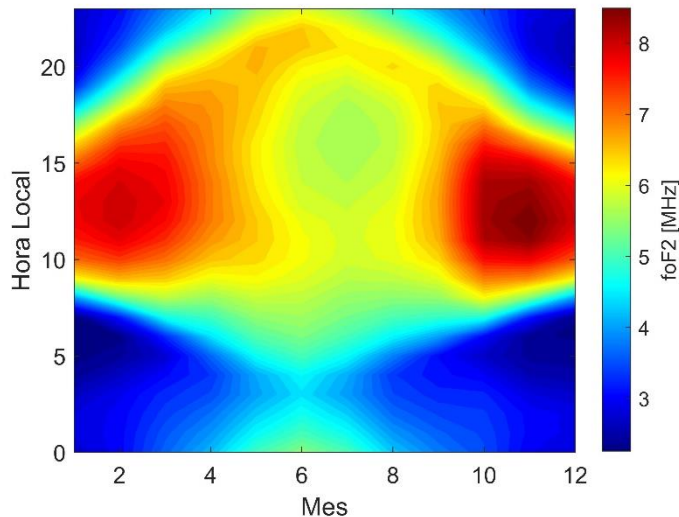


Figura 4.5 Promedio de foF2 [MHz] en el período 1960-2023, considerando el valor de medianas mensuales por hora local y mes.

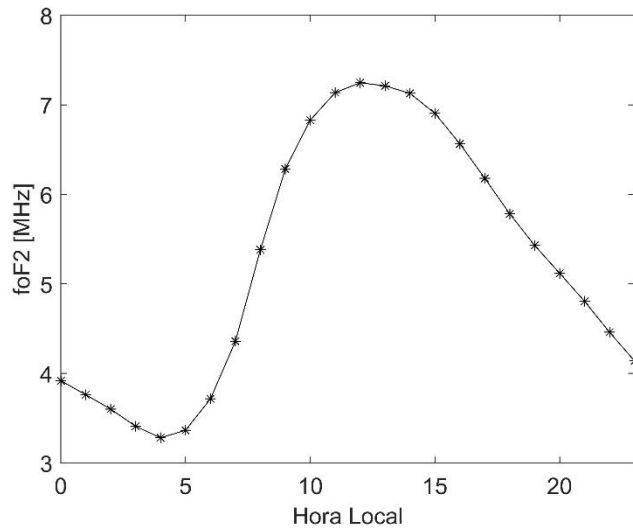


Figura 4.6 Promedio de foF2 [MHz] para cada hora estimado a partir de los valores de medianas mensuales en el período 1960-2023.

4.3 Generación de huecos

Para generar huecos en la serie se utilizaron dos criterios: extrayendo datos consecutivos y extrayendo datos en posiciones aleatorias.

En el primer caso, se eligieron 7 años consecutivos (desde septiembre 1988 hasta octubre 1995) y se eliminaron los valores correspondientes de todas las horas. Es decir que se generaron 2064 valores en la serie original que cuenta completa con 18432 ($64 \times 12 \times 24$) valores, que corresponden a un 11.2% del total de los valores. Las Figuras 4.7 y 4.8 muestran la serie resultante luego de generar los huecos.

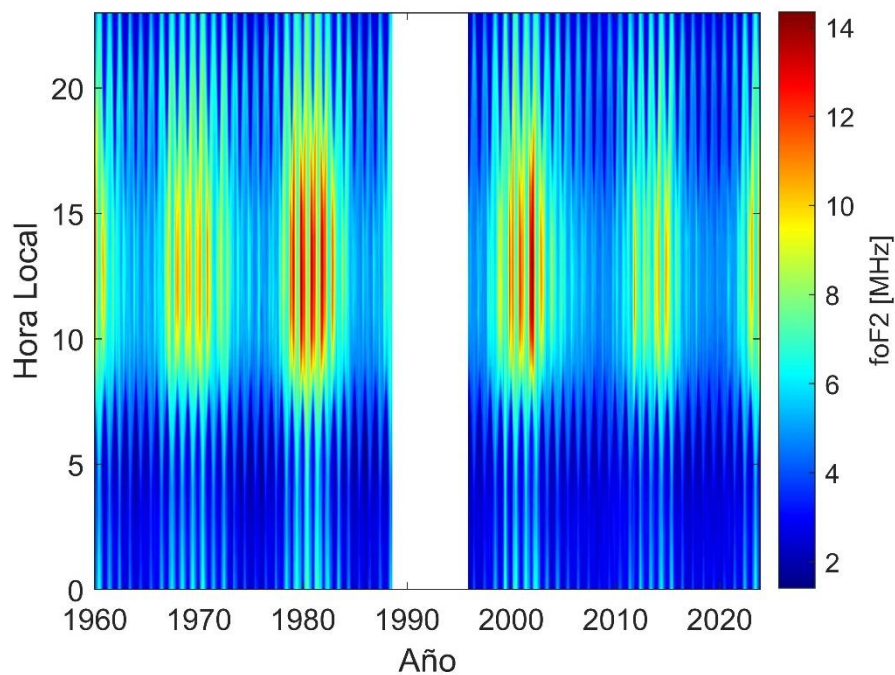


Figura 4.7 Mediana mensual de foF2 [MHz] medido en Juliusruh en el período 1960-2023, con datos faltantes generados en el período septiembre 1988-octubre 1995 para todas las horas.

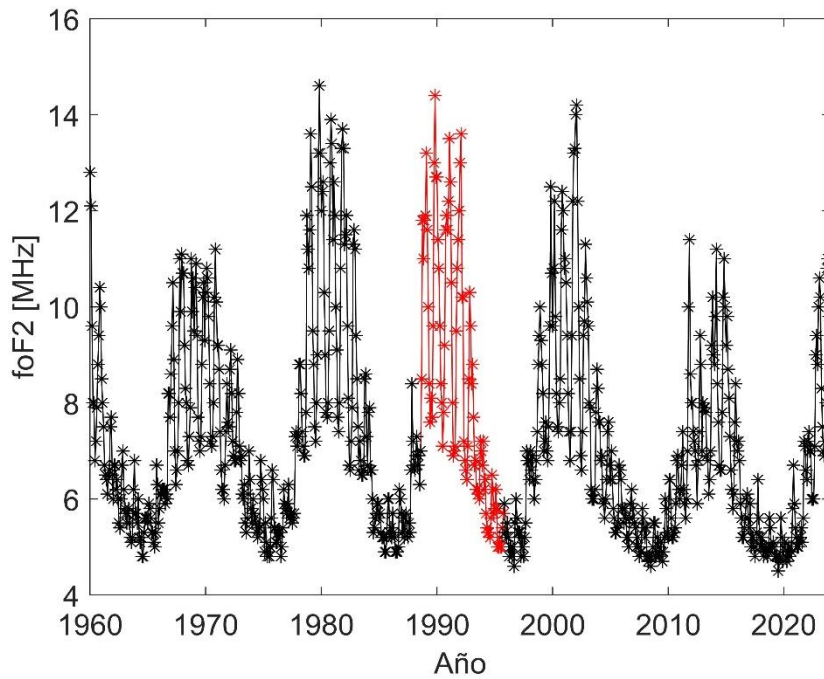


Figura 4.8 Mediana mensual de foF2 [MHz], para la hora 12 LT, medido en Juliusruh en el período 1960-2023. La línea y símbolos rojos indican los datos de foF2 extraídos para generar huecos consecutivos entre septiembre 1988-octubre 1995.

Y se consideró también la generación de mayor cantidad de huecos en dos sub-períodos de tiempo diferentes: 1975 a 1985 y 2010 a 2020 (31.2% del total de los datos), como se muestra en Figura 4.9.

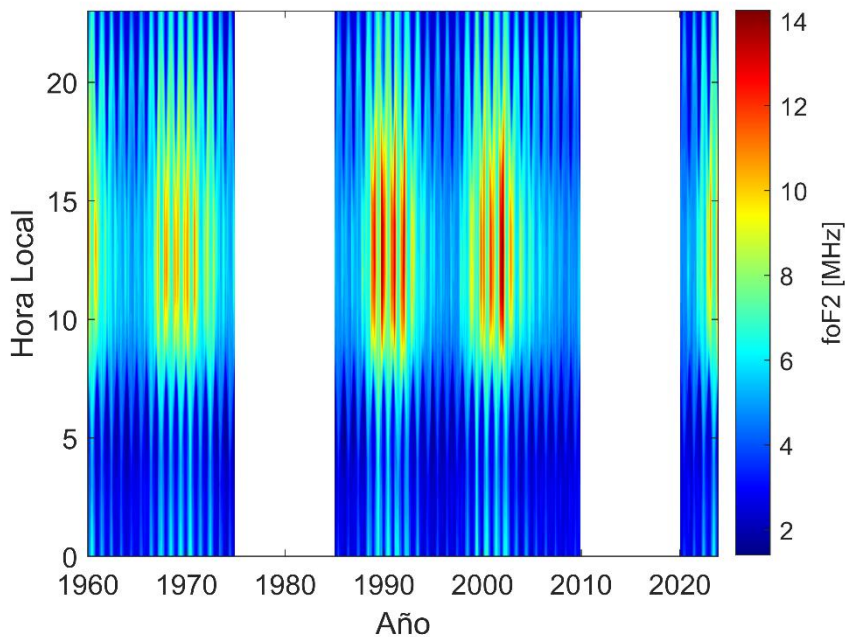


Figura 4.9 Mediana mensual de foF2 [MHz] medido en Juliusruh en el período 1960-2023, con datos faltantes generados en el período 1975-1985 y 2010-2020 para todas las horas y meses.

El segundo criterio fue extraer valores de la serie de manera aleatoria. Se generaron archivos donde se extrajeron el 60% y 70% de los datos de foF2, en el caso de completado de datos faltantes con el modelo de regresión lineal, y hasta 97% para los métodos basados en ML.

En el caso de métodos de regresión lineal, la generación de porcentajes de huecos superiores al 70% no resulta conveniente, ya que la ubicación aleatoria de los datos faltantes puede ocasionar que, para algunas combinaciones de mes y hora, no se disponga del número mínimo de observaciones requerido para estimar adecuadamente los coeficientes de regresión.

En la figura 4.10 se muestra la gráfica del 70% de los datos extraídos de manera aleatoria.

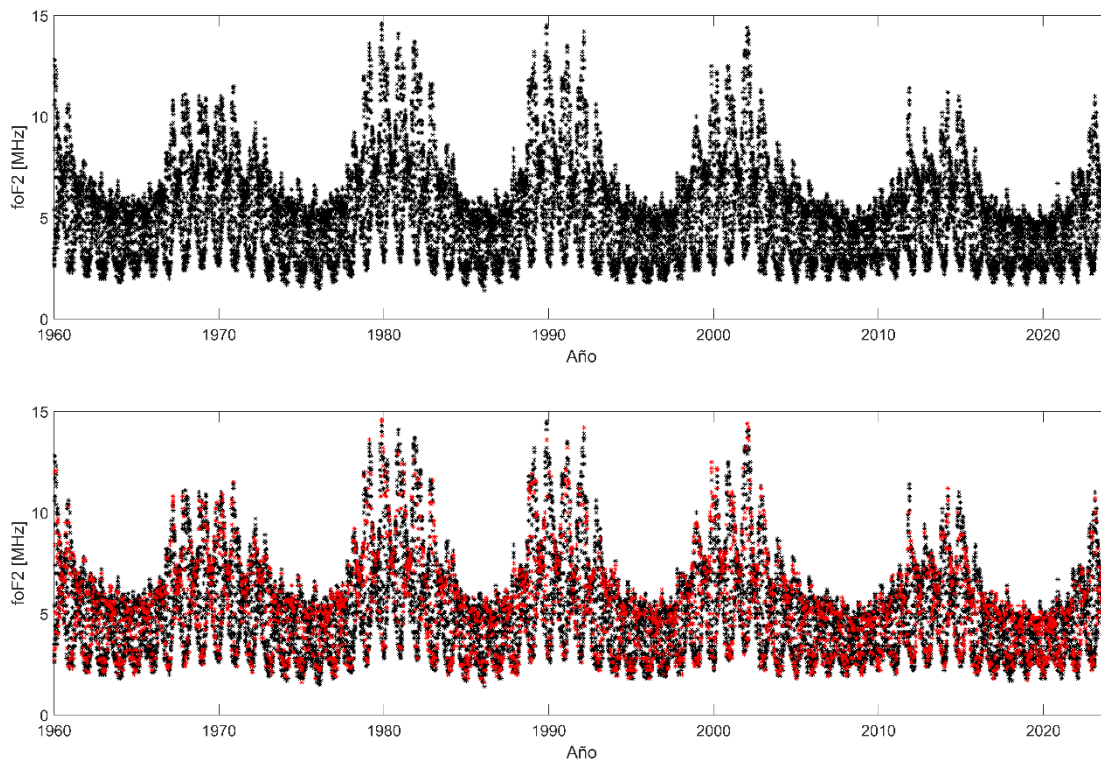


Figura 4.10 Mediana mensual de foF2 [MHz], medido en Juliusruh en el período 1960-2023, considerando las 24 horas del día para todos los años y meses. Panel superior: serie completa (puntos negros, total=18432). Panel inferior: serie con 70% de datos faltantes (puntos negros, total=4608) y datos extraídos de forma aleatoria (puntos rojos=13824)

4.4 Evaluación del desempeño de los modelos

Para evaluar el rendimiento de los distintos métodos aplicados a la estimación de datos faltantes en series temporales de foF2, se emplean tres métricas estadísticas ampliamente utilizadas en este caso: el coeficiente de determinación (R^2), el Error Absoluto Medio (MAE, Mean Absolute Error) y el Error Absoluto Medio Porcentual (MAPE, Mean Absolute Percentage Error) (Gueymard, 2014; Medina et al., 2023). Estas métricas permiten cuantificar la capacidad predictiva de los modelos y comparar su desempeño de manera objetiva.

Los estadísticos de comparación se estimaron, no de la manera tradicional, en la cual uno estima estos valores a partir de la predicción del modelo relativo a la observación que se utilizó para

estimar el modelo. En este trabajo estos estadísticos se estimaron considerando solamente la predicción de los modelos para el caso de los datos faltantes que no se utilizaron para la estimación del modelo relativos a las observaciones faltantes de las cuales sí se dispone, solo que no se las utilizó. Se tomó como valor estimado, las interpolaciones de cada método, y como valor de referencia se consideró foF2 observado (o medido) que se extrajo a propósito precisamente para poder evaluar el desempeño de los métodos de interpolación. Se procedió de manera similar en los casos donde los datos fueron extraídos aleatoriamente.

4.4.1 Coeficiente de determinación (R^2)

El coeficiente de determinación (R^2) cuantifica la capacidad del modelo para reproducir la variabilidad observada en los datos experimentales. Un valor de $R^2=1$ corresponde a una concordancia perfecta entre los valores observados y estimados por el modelo. Por el contrario, un valor de $R^2=0$ indica que el modelo no logra explicar una fracción de la variabilidad mayor que la obtenida utilizando únicamente la media de las observaciones. Valores negativos de R^2 sugieren que el desempeño del modelo es inferior al de una estimación basada exclusivamente en dicha media.

La expresión matemática utilizada es:

$$r^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (4.1)$$

donde y_i representa los valores originales de la serie, $\hat{y}_i$ el valores predichos por el modelo de interpolación, y $\bar{y}$ el valor promedio de la serie original. La sumatoria incluye solamente el período de datos donde se generaron los huecos y que luego fueron interpolados. Es decir que la comparación se hace entre los datos interpolados con los diferentes métodos y los valores originales que se eliminaron para generar estos huecos.

4.4.2 Error Absoluto Medio (MAE)

El MAE cuantifica la magnitud promedio de los errores cometidos por el modelo de interpolación, sin considerar su signo. Es decir, mide el promedio de las diferencias absolutas entre los valores reales y los predichos. Es una métrica de fácil interpretación, siempre positiva por definición, y su valor óptimo es cero, lo que indicaría una predicción perfecta. En este trabajo, el MAE proporciona una medida directa del error típico que comete el modelo en sus estimaciones, y se calcula como

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i| \quad (4.2)$$

La sumatoria incluye solamente el período de datos donde se generaron los huecos y que luego fueron interpolados. n corresponde al número de datos interpolados.

4.4.3 Error Absoluto Medio Porcentual (MAPE)

El MAPE evalúa lo mismo que el MAE pero en términos porcentuales. Su expresión es

$$MAPE = \frac{1}{n} \sum \frac{|y_i - \hat{y}_i|}{y_i} 100 \quad (4.3)$$

Nuevamente, la sumatoria incluye solamente el período de datos donde se generaron los huecos y que luego fueron interpolados.

Capítulo 5. Resultados

En este capítulo se muestran los resultados de la interpolación de foF2 medido en Juliusruh aplicando los diferentes métodos seleccionados y se realiza el análisis comparativo del desempeño de cada uno de ellos en función de los estadísticos seleccionados para este propósito.

5.1 Interpolación de foF2 utilizando regresión lineal

En escala interanual, foF2 responde principalmente a la actividad solar. Los índices de actividad solar no presentan la variación diaria ni la estacional que presenta foF2. Una manera de tener esto en cuenta, es separar la serie original de foF2 en series para cada hora y mes por separado. De esta manera se tienen 288 series (12 meses $\times$ 24 horas). Para cada una de estas series, con huecos, se estima el modelo de regresión lineal utilizando cada uno de los distintos proxies de actividad solar, con y sin el agregado de Ap como índice de actividad geomagnética, y también agregando el término cuadrático en el proxy solar. Se agrega incluso, en diversas publicaciones que utilizan este tipo de modelado, el término cúbico (Liu et al., 2004; Elias, 2014; Chen et al., 2022; Tan Jun Rios et al., 2025). Es decir que, en este caso, se analizarán los siguientes modelos de regresión lineal:

$$\text{foF2} = A_0 + A_1 X \quad (5.1)$$

$$\text{foF2} = A_0 + A_1 X + B \text{ Ap} \quad (5.2)$$

$$\text{foF2} = A_0 + A_1 X + A_2 X^2 \quad (5.3)$$

$$\text{foF2} = A_0 + A_1 X + A_2 X^2 + B \text{ Ap} \quad (5.4)$$

donde X corresponde a cada uno de los cuatro proxies de actividad solar seleccionados para este trabajo: SN, MgII, F10.7, y F30.

Luego, a partir de cada modelo, donde se estiman los parámetros de regresión utilizando mínimos cuadrados, se calculan los datos faltantes utilizando el proxy del año y mes correspondiente, para así completar cada serie. Esto se realizó, como ya se mencionó, para cada mes y para cada hora por separado. Es decir que se obtuvieron $24 \times 12 = 288$ modelos de regresión lineal para cada uno de los cuatro casos caso, y para cada proxy.

El modelo fue puesto a prueba usando las series de datos con huecos ya mencionados, es decir:

- (1) con faltantes de datos continuos durante 7 años, desde 9/1988 hasta 10/1995 (~10% del total de los datos)
- (2) con dos periodos continuos de faltantes, desde 1/1975 hasta 1/1985 y 1/2010 a 1/2020 (~30% del total de los datos)
- (3) con 60% y 70% de huecos generados de manera aleatoria.

Comenzando con la serie (1), la Figura 5.1 muestra los valores de MAE en función de las horas y el mes para cada uno de los 4 modelos de interpolación correspondientes al proxy solar F10.7. La Figura 5.2 muestra los valores de MAPE y la Figura 5.3 muestra la evolución temporal del error absoluto para cada año/mes/hora, sin realizar ningún promedio. La Figura 5.4 muestra los valores de foF2 en función de las horas y el mes/año sólo para el período interpolado.

Cabe destacar que los valores presentados en la Figura 5.1 se obtienen al promediar, para cada mes y hora, los errores absolutos mostrados en la Figura 5.3 a lo largo de todo el período donde se generaron los huecos.

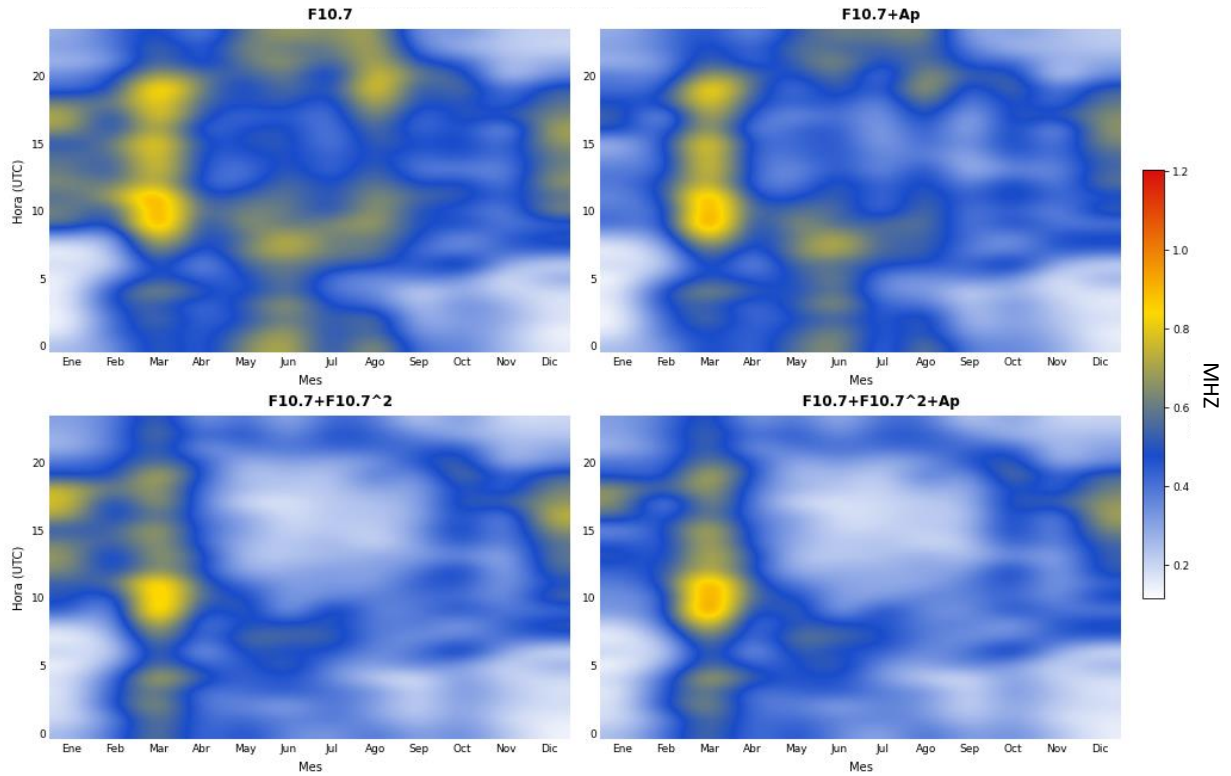


Figura 5.1 MAE por hora y mes correspondiente a los modelos de interpolación de las Ecuaciones 5.1 a 5.4 para el proxy solar F10.7 con dataset (1).

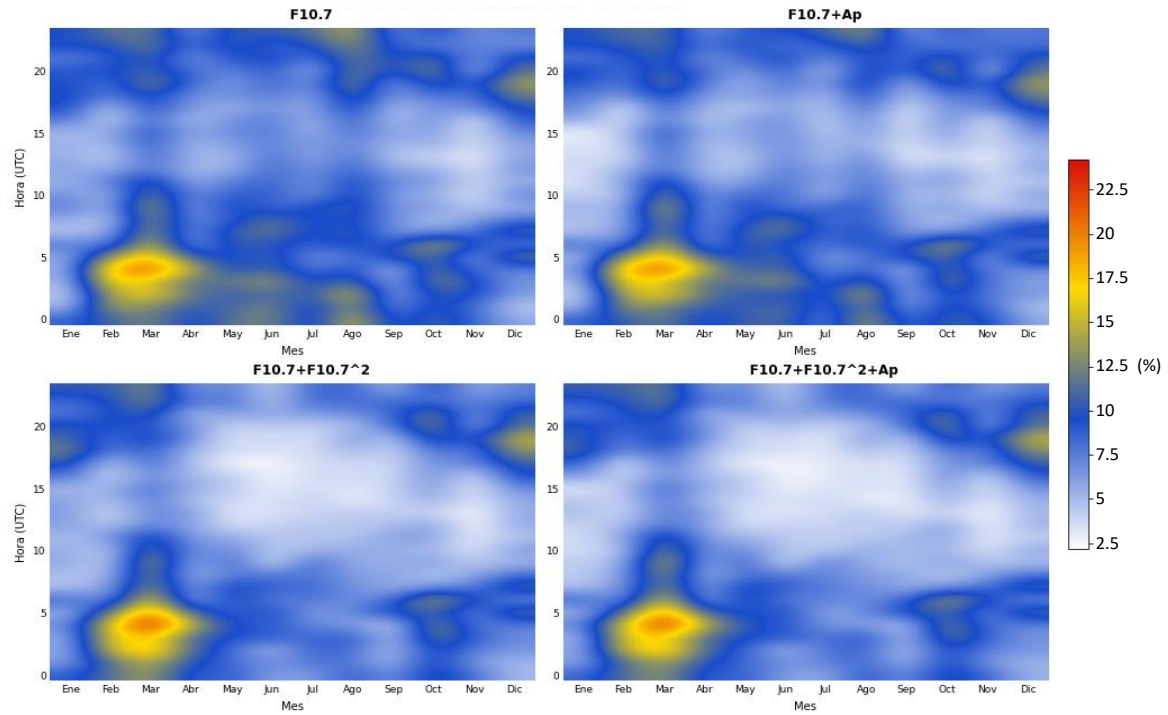


Figura 5.2 MAPE (%) de los modelos de interpolación correspondiente a las Ecuaciones 5.1 a 5.4 para el proxy solar F10.7 correspondiente al dataset (1).

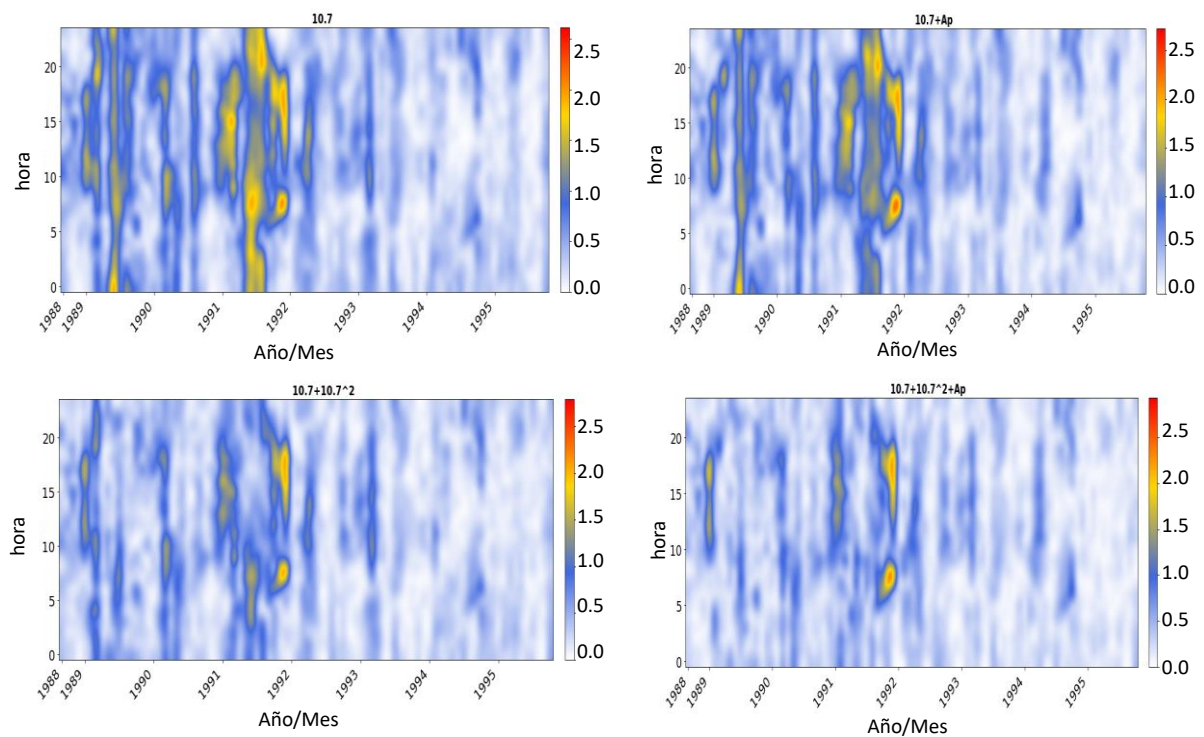


Figura 5.3 Error absoluto (MHz) para cada año-mes-hora

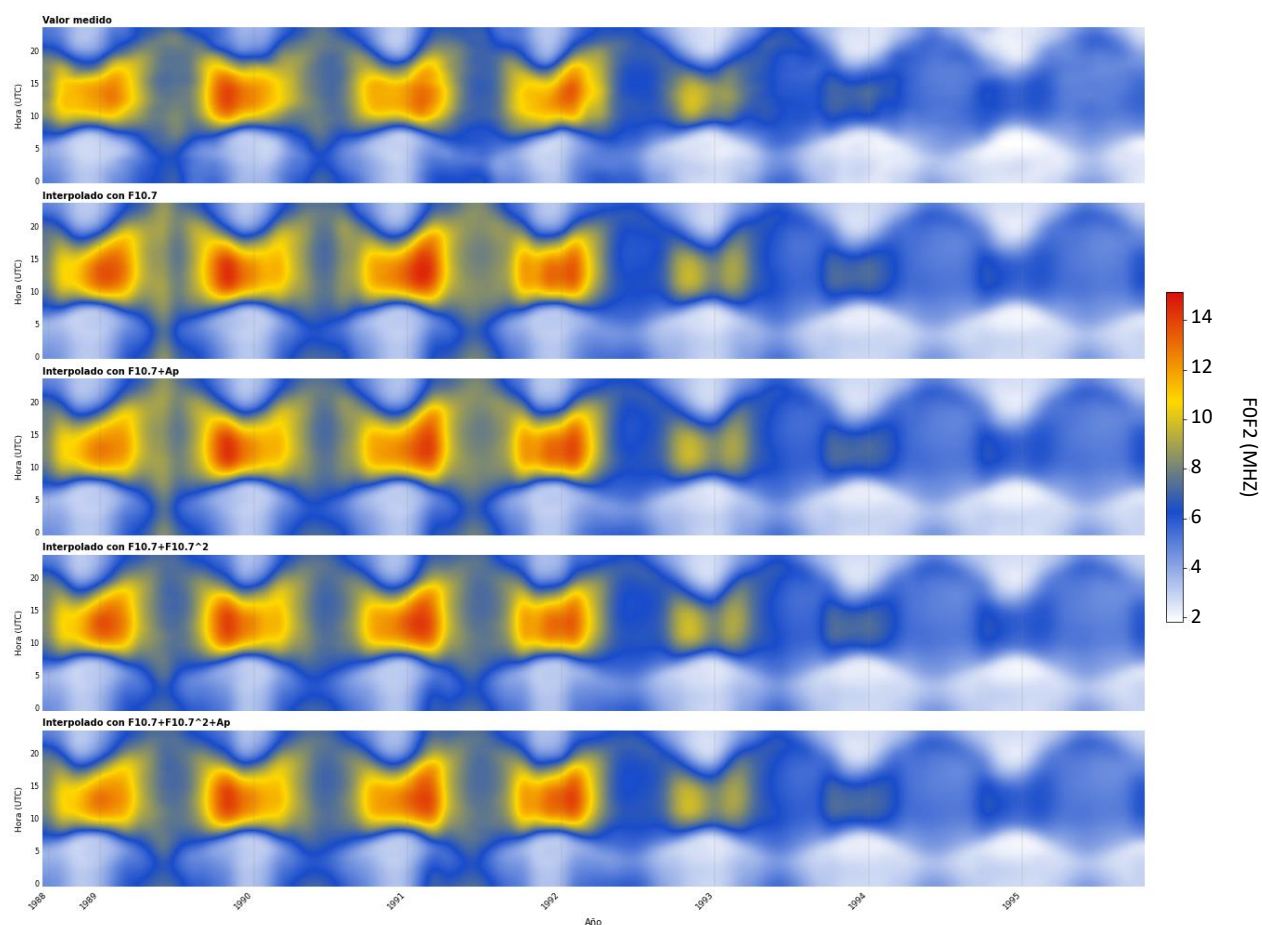


Figura 5.4 Valores medidos que fueron eliminados para generar los huecos (panel superior) y valores interpolados considerando cada uno de los modelos de regresión lineal dados por las Ecuaciones 5.1 a 5.4.

Los valores de MAE, MAPE y R^2 promediados sobre todas las horas y meses se muestran en la Tabla 1.

Tabla 1. Valores de R^2 MAE y MAPE para los cuatro modelos de regresión lineal que utilizan F10.7 como proxy solar, promediados en todas las horas y meses. Se sombrea en amarillo el modelo con la mejor performance.

	R^2	MAE	MAPE
F10.7	0.93	0.48	8.44%
F10.7 + Ap	0.94	0.45	7.62%
F10.7 + F10.7 ²	0.95	0.40	7.22%
F10.7 + F10.7 ² + Ap	0.96	0.35	6.24%

Se repitió el mismo procedimiento con los proxies solares MgII, F30 y SN. Los resultados de los índices estadísticos promediados sobre el período y todas las horas y meses,

correspondientes a cada uno se muestran en la Tabla 2. Las Figuras correspondientes a los valores de MAE, MAPE y errores y valores predichos (Figuras 5.1 a 5.4 considerando F10.7), son muy similares a las que resultan de utilizar F10.7, de manera que no se repiten en este trabajo de Tesis.

Tabla 2. Valores de R^2 , MAE y MAPE promediados para todos los meses y horas del día, para cada uno de los modelos de regresión lineal y para cada uno de los proxies solares considerados. Se sombrea en amarillo el modelo con la mejor performance.

	R^2	MAE	MAPE
F10.7	0.93	0.48	8.44%
F30	0.94	0.47	8.58%
MgII	0.96	0.40	7.28%
SN	0.94	0.48	8.81%
F10.7 + Ap	0.94	0.45	7.62%
F30+Ap	0.94	0.44	7.69%
MgII+Ap	0.96	0.39	6.77%
SN+Ap	0.94	0.47	8.20%
F10.7 + F10.7 ²	0.95	0.40	7.22%
F30+ F30 ²	0.94	0.43	7.92%
MgII+MgII ²	0.96	0.36	6.58%
SN+SN ²	0.95	0.43	7.95%
F10.7 + F10.7 ² + Ap	0.96	0.35	6.24%
F30+ F30 ² +Ap	0.95	0.39	6.90%
MgII+MgII ² +Ap	0.96	0.34	6.04%
SN+SN ² +Ap	0.95	0.41	7.28%

En base a los índices estadísticos obtenidos se observa que el mejor desempeño se obtiene utilizando el proxy solar MgII con la Ecuación (5.4) por lo tanto será el que se utilizará para hacer los siguientes análisis.

En la Figura 5.5 se muestra la serie completa de foF2 en función del mes/año y hora junto con el período interpolado señalado entre líneas rojas verticales, utilizando MgII como proxy solar y el modelo dado por la Ecuación (5.4).

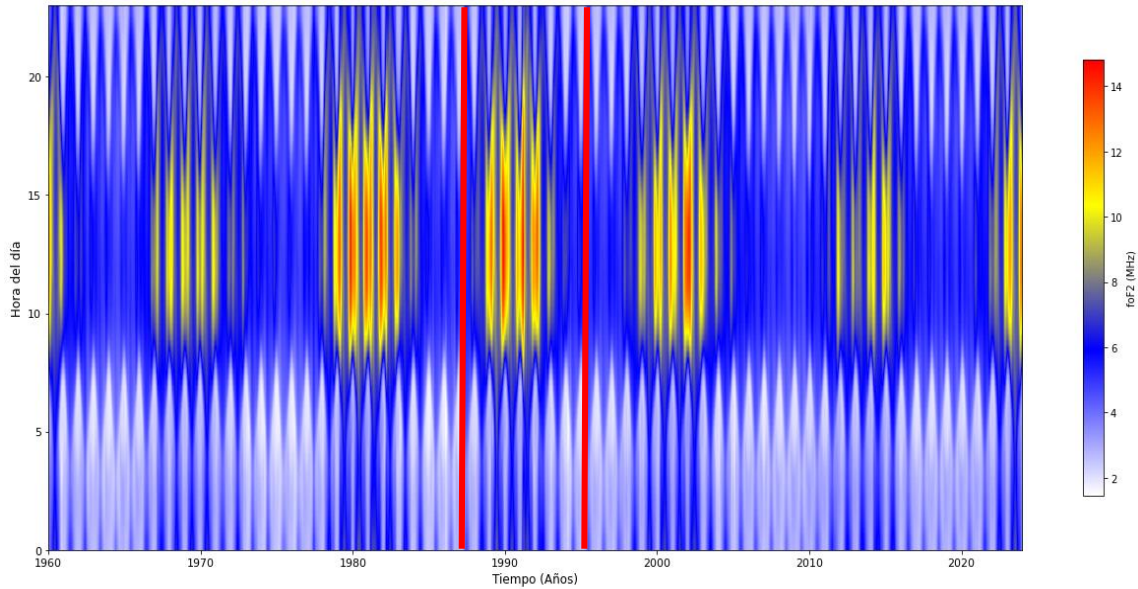


Figura 5.5 foF2 [MHz] en función del año y hora local. Las líneas rojas verticales indican el período de datos interpolados que resultan de utilizar el modelo de regresión lineal dado por la Ecuación 5.4, y MgII como proxy solar: $foF2 = A_0 + A_1 \text{MgII} + A_2 \text{MgII}^2 + B \text{Ap}$. Los períodos a ambos lados de las líneas rojas corresponden a los datos originales.

En el caso de la serie (2) (donde de tener ~10% de huecos se pasa a tener ~30% de huecos), considerando el modelo de regresión dado por la Ecuación (5.4) y MgII como proxy solar, los valores promedios totales de los estadísticos comparativos, resultan similares a los obtenidos en el caso del dataset (1) (que se listan en la Tabla 1). Estos valores resultan en este caso: $R^2=0.96$, $MAE=0.34$ y $MAPE=6.77\%$.

La Figura 5.6, de manera similar a la 5.3 muestra la serie de tiempo de los errores absolutos.

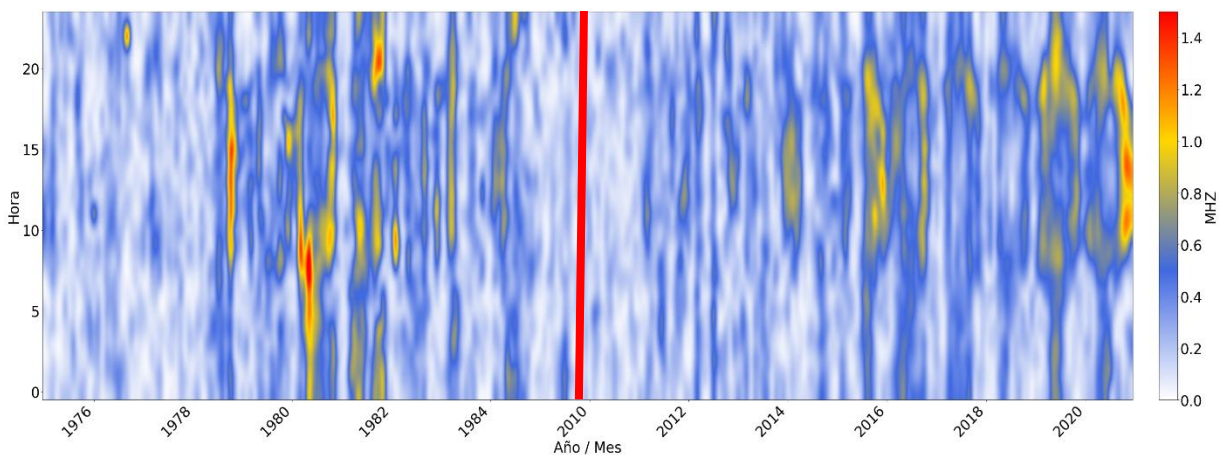


Figura 5.6 Gráfica de error absoluto para los periodos 1975/85 y 2010/20.

La Figura 5.7, de forma similar a la Figura 5.4, muestra la serie original de foF2 con los dos subperíodos donde se generaron huecos (1975 a 1985, y 2010 a 2020) en los cuales los valores corresponden a los obtenidos con el modelo de regresión utilizando MgII.

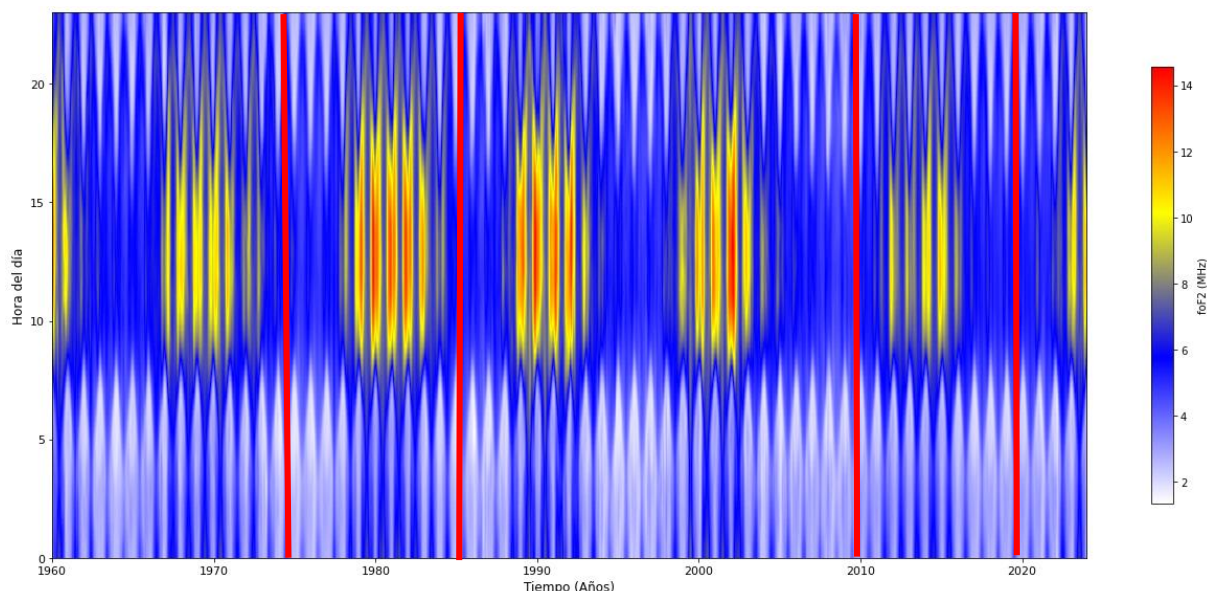


Figura 5.7 foF2 [MHz] en función del año y hora local. Las líneas rojas verticales indican los períodos de datos interpolados (1975 a 1985, 2010 a 2020) y que resultan de utilizar el modelo de regresión lineal dado por la Ecuación 5.4, y MgII como proxy solar: $foF2 = A_0 + A_1 \text{MgII} + A_2 \text{MgII}^2 + B \text{Ap}$.

En el caso de datos faltantes aleatorios, es decir la serie (3), los estadísticos comparativos resultan también similares a los casos anteriores. Aquí se agrega algo para considerar. En períodos de máxima y mínima actividad solar, es cuando más discrepa, en general, el valor real de foF2 del valor modelado en función de los proxies. Entonces, al generar datos de manera aleatoria, se espera que haya casos que resulten con mayor errores promedio que otros. Para tener esto en cuenta, se generaron 1000 series con 60% y 70% de huecos. Los valores promedio sobre estas 1000 repeticiones se listan en la Tabla 3 junto con la desviación estándar de cada uno. En las Figuras 5.8 y 5.9 se muestran los histogramas de estos estadísticos para tener una idea además de sus rangos de variación. Se observa que, si bien la desviación estándar es de menos del 1% en todos los casos, lo que implica que con repetición se obtiene el valor esperado, el valor máximo que puede alcanzar el error sí varía significativamente: de ~100% a ~300% cuando faltan 60% y 70% de los datos respectivamente. De todas maneras, esto no se analizará en este trabajo, y para el resto de los métodos de interpolación solo se analizará una repetición de huecos aleatorios. La Tabla 3 lista además los valores de los estadísticos que resultan de considerar las series (1) y (2).

Tabla 3. Valores de R^2 , MAE y MAPE promediados para todos los meses y horas del día, para

cada uno de los modelos de regresión lineal y para cada uno de los proxies solares considerados.

Método de generación de huecos	60% aleatorios	70% aleatorios	~10% (1988-1995)	~30% (1975-1985 y 2010-2020)
R^2	0.956±0.001	0.950±0.003	0.966	0.956
MAE (MHz)	0.342±0.003	0.358±0.004	0.345	0.341
MAPE (%)	6.78±0.05	7.10±0.07	6.04	6.77

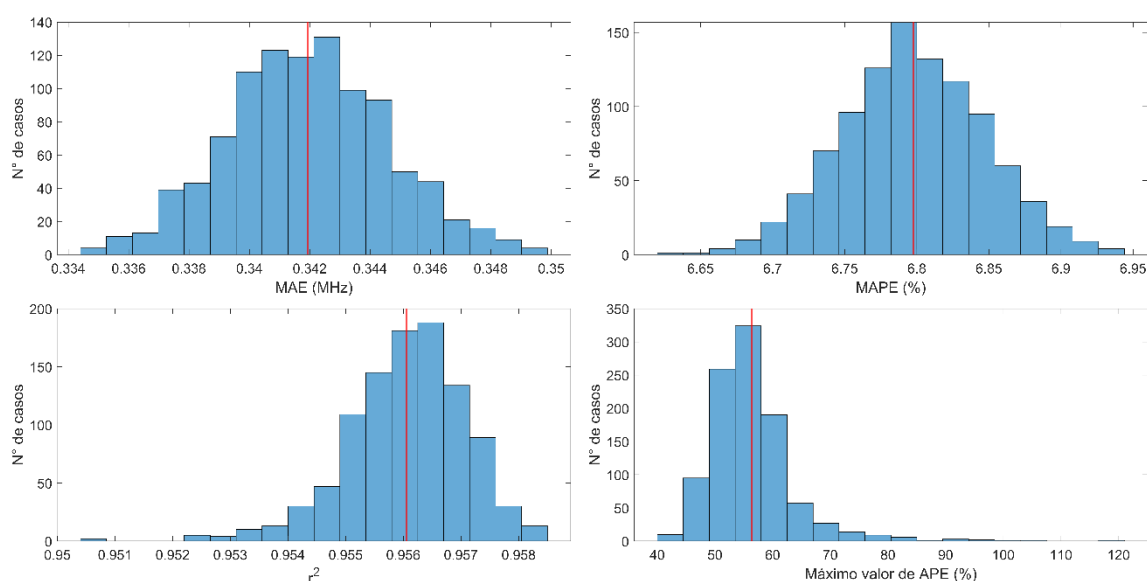


Figura 5.8 Histogramas de los valores de MAE, MAPE, R^2 y máximo valor del error absoluto porcentual (APE) obtenido para la serie de foF2 en las que se generaron 60% de datos faltantes de manera aleatoria con 1000 repeticiones.

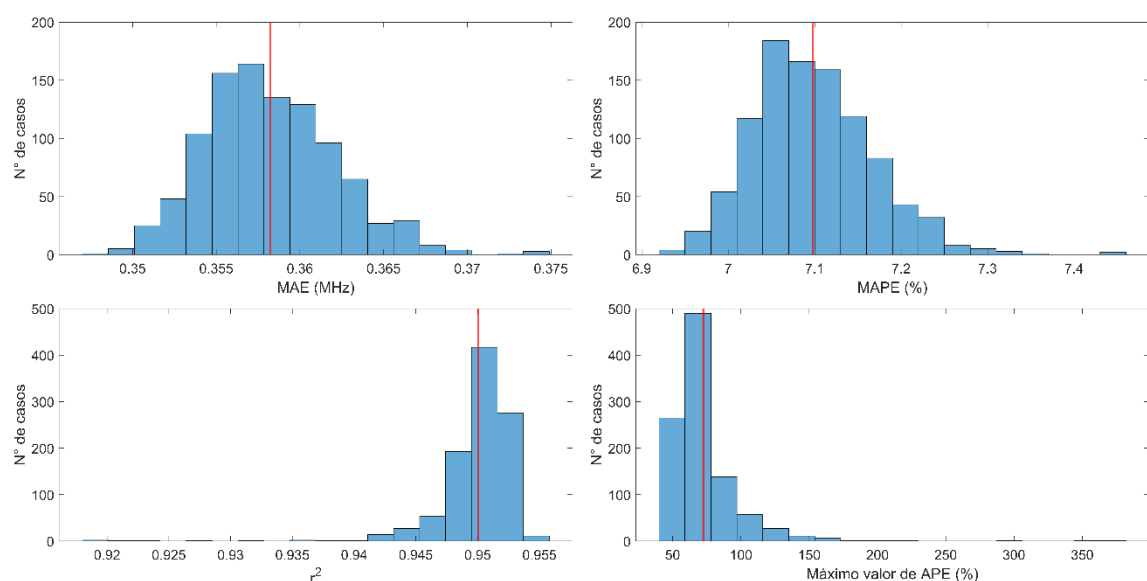


Figura 5.9 Histogramas de los valores de MAE, MAPE, R^2 y máximo valor del error absoluto porcentual (APE) obtenido para la serie de foF2 en las que se generaron 70% de datos faltantes de manera aleatoria con 1000 repeticiones.

Se puede observar además que los valores de MAE, MAPE y R^2 son muy parecidos en todos los casos de cantidad de huecos, desde 10% hasta 70%. La diferencia se observa, como se mencionó, en el valor del máximo valor del error absoluto, que cada vez es mayor, y se podría agregar el caso del mínimo de valor de R^2 , que a medida que aumenta la cantidad de huecos, disminuye. En nuestro caso se observe que al ir de 60% a 70% de huecos el mínimo valor de R^2 disminuye de 0.95 a 0.92 que corresponde a una diferencia de casi 10%.

5.2 Interpolación de foF2 utilizando Bosques Aleatorios (Random Forest Regression, RFR)

En este caso, ya no separamos la serie original de foF2 en series individuales por hora y mes (lo que resultaba en 288 series con 64 datos totales cada una), sino que consideramos la serie completa como una única serie de tiempo con 18432 datos totales en el caso de serie completa. Esto ya es una ventaja, ya que implica una primera simplificación en este proceso de completado de datos faltantes.

Comenzando con el caso de la serie (1), este modelo de ML fue entrenado utilizando las siguientes variables predictoras: F10.7, MgII, Rz, Ap, F30 y las funciones seno y coseno de la hora del día y del mes del año. Se utilizó la función RandomForestRegressor de la librería scikit-learn configurada con 500 árboles. Semilla random state igual a 42 y n_jobs igual a -1 para utilizar todos los núcleos del procesador.

Los datos se separaron en dos conjuntos: el conjunto de entrenamiento, conformado por todos los registros donde foF2 se encontraba disponible y el conjunto de predicción, integrado por los registros con valores faltantes de foF2. El modelo fue entrenado con el conjunto completo de datos disponibles y posteriormente se aplicó para predecir los valores faltantes de foF2, para cada uno de los casos de series en las que se generaron distinta cantidad de huecos. La Figura 5.10 muestra la serie completa con los datos interpolados con este método de ML en el caso de la serie (1).

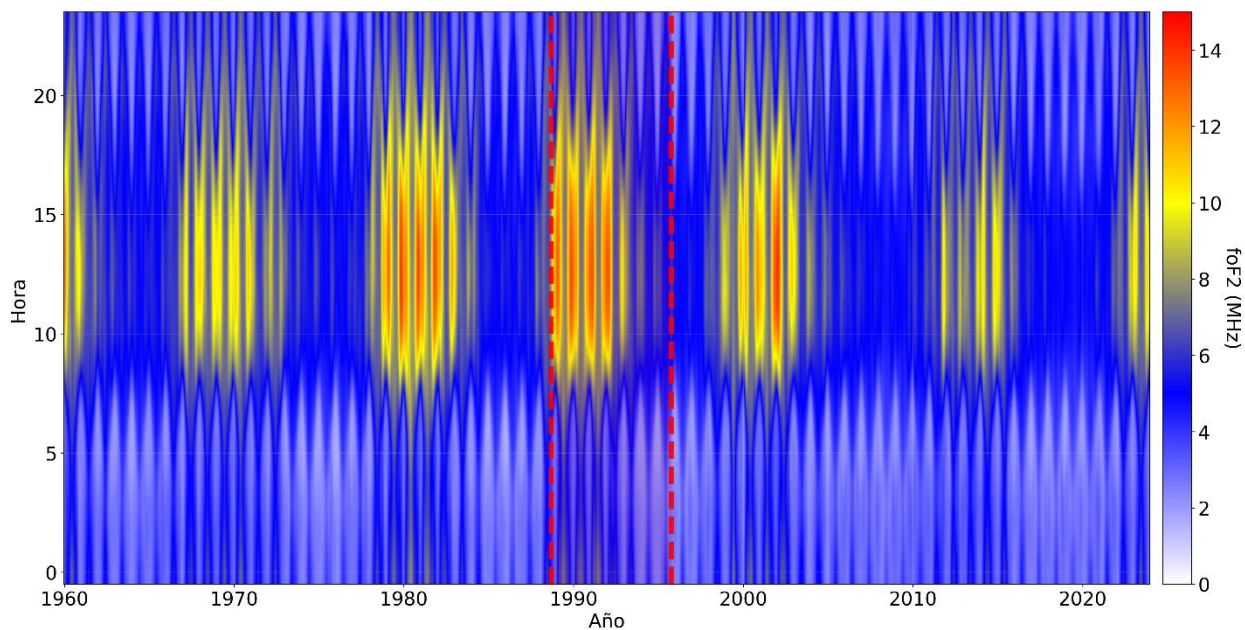


Figura 5.10 Período completo de foF2 para dataset (1). El periodo comprendido entre las dos líneas rojas verticales corresponde al periodo interpolado.

En la Figura 5.11 se muestran gráficas de calor del MAE y MAPE para cada hora y mes para poder comparar con los valores obtenidos en el caso del método de regresiones lineales que se mostraron en las Figuras 5.1 y 5.2. La Figura 5.12 muestra los valores del error absoluto para cada hora y mes/año.

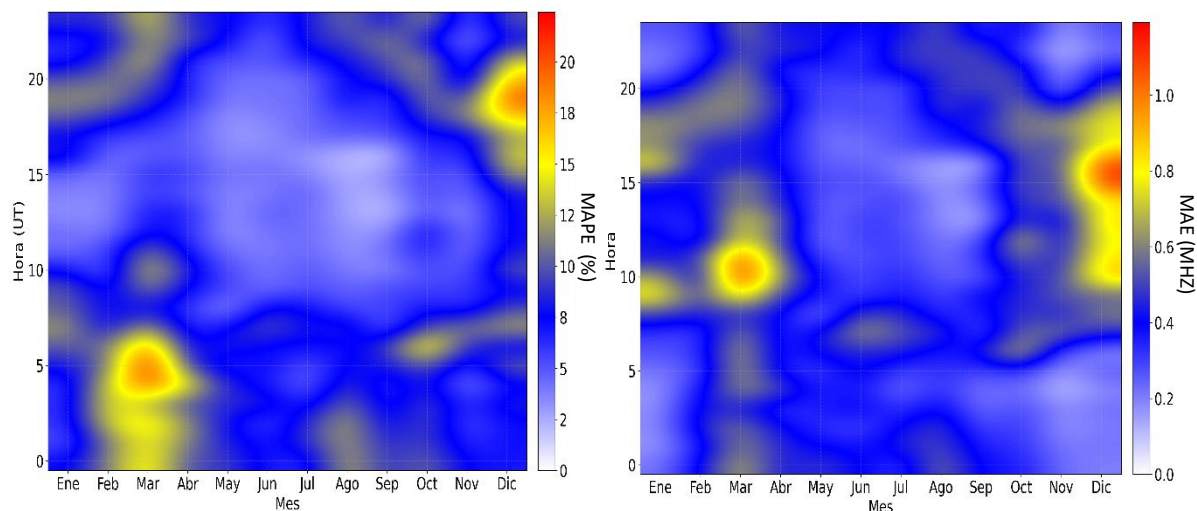


Figura 5.11 MAE y MAPE por hora y mes correspondiente a la técnica de RFR para la interpolación en el caso de la serie (1).

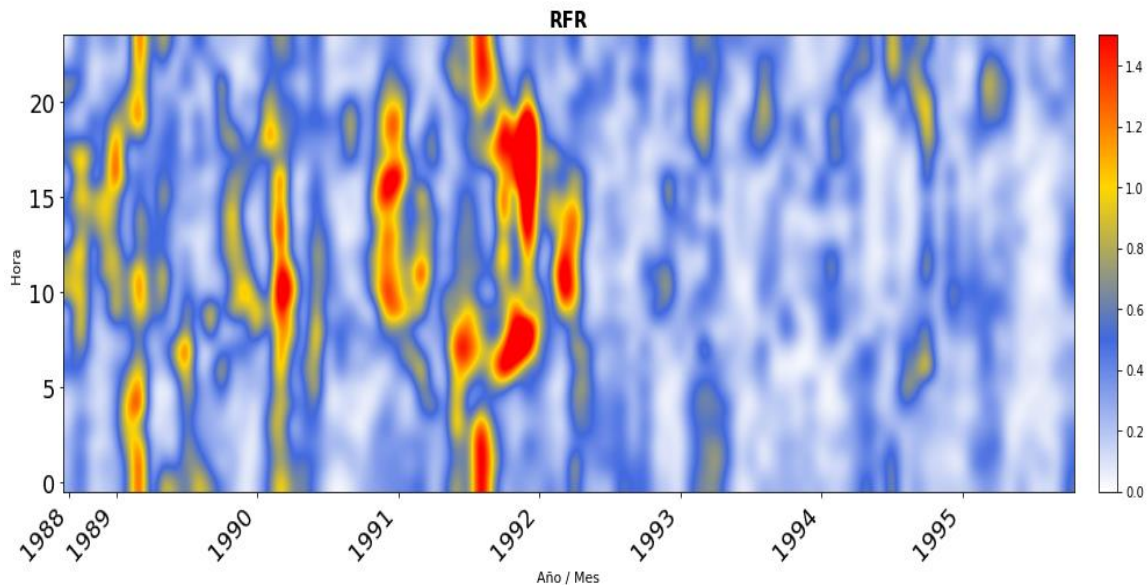


Figura 5.12 Valores de error absoluto (MHz) para cada año-mes-hora del periodo interpolado

Los valores promedios totales de los estadísticos, es decir promediando todas las horas y meses resultan similares a los obtenidos usando los modelos de regresión. En este caso estos valores son: 0.409 MHz para MAE, 7.44 % para MAPE y 0.951 para R^2 . Estos mismos valores se listan en la Tabla 7 más adelante, junto con los valores obtenidos usando los otros métodos de interpolación.

Valores y figuras similares a las anteriores se obtienen en el caso de la serie (2) interpolando los valores faltantes con RFR. Lo mismo ocurre con la serie (3) donde se generan 60% y 70% de huecos aleatorios.

En el caso de los métodos basados en ML, la limitación en el mínimo número de datos necesario en las regresiones lineales para obtener los coeficientes con el método de mínimos cuadrados, no es evidente y los métodos ya no arrojan error. Un motivo es que se considera la serie total de datos sin separarla por hora y mes, lo que elimina el riesgo de tener alguna serie sin datos, por ejemplo. Sin embargo, hay que tener en cuenta que contar con pocos datos para el entrenamiento aumenta el riesgo de sobreajuste y la capacidad de generalización empeora.

Aun así, se llevó al límite la cantidad de huecos, utilizando 90% y 97% de huecos generados de forma aleatoria. Los valores de los estadísticos de comparación se listan en la Tabla 4. Se observa que, a pesar de seguir siendo valores aceptables, sobre todo en el caso de 97% de huecos el valor de R^2 cayó a casi 0.9 y el valor de MAPE llegó a 10%.

Tabla 4 Valor de MAE, MAPE y R^2 para la serie de foF2 donde se generaron 90% y 97% de huecos que fueron luego interpolados con RFR.

	90% de huecos aleatorios	97% de huecos aleatorios

R^2	0.938	0.904
MAE (MHz)	0.386	0.486
MAPE (%)	7.76	10.00

5.3 Interpolación de foF2 utilizando Gradient Boosting Regression (GBR)

Al igual que en el caso de RFR, en el caso de GBR para la imputación de datos faltantes, se considera la serie completa como una única serie de tiempo con 18432 datos en total cuando la serie de foF2 está completa.

Este modelo fue entrenado utilizando las mismas variables predictoras que en el caso de RFR, es decir: F10.7, MgII, Rz, Ap, F30 y las funciones seno y coseno de día y meses. Se configuró con 400 estimadores, una tasa de aprendizaje de 0.05, una profundidad máxima de 6 por árbol, un subsample de 0.8 y semilla aleatoria igual a 42.

En la Figura 5.13 se muestran graficas de calor del MAE y MAPE para cada hora y mes para poder comparar con los valores obtenidos en el caso del método de regresiones lineales que se mostraron en las Figuras 5.1 y 5.2. La Figura 5.14 muestra los valores del error absoluto para cada hora y mes/año. Se observa en todos los casos una mejora con respecto al método de RFR (ver Figuras 5.11y 5.12).

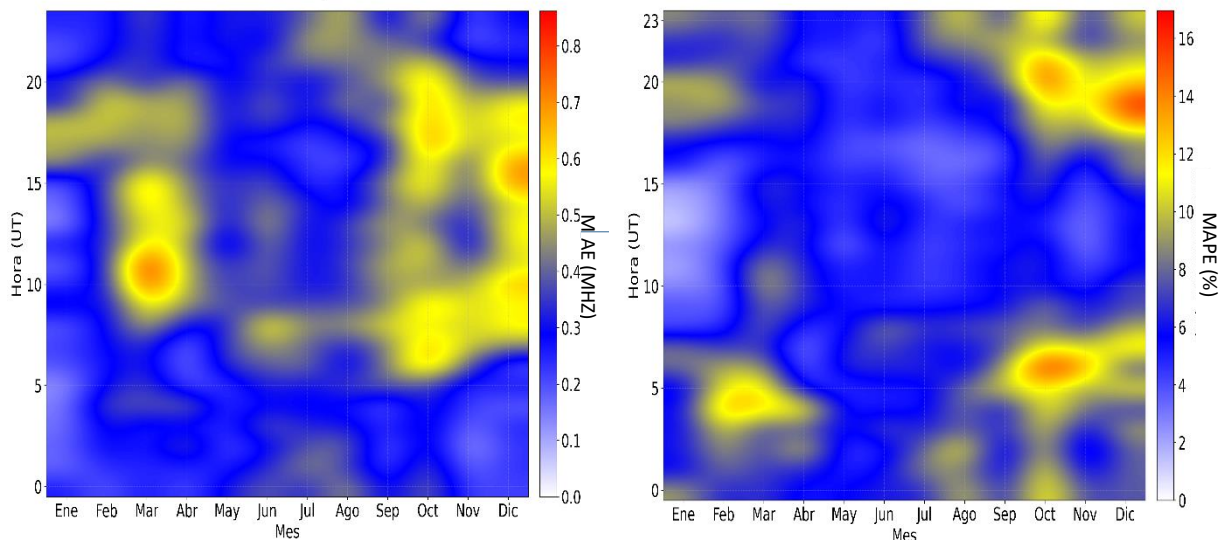


Figura 5.13 MAE y MAPE por hora y mes correspondiente a la técnica de GBR para la interpolación en el caso de la serie (1).

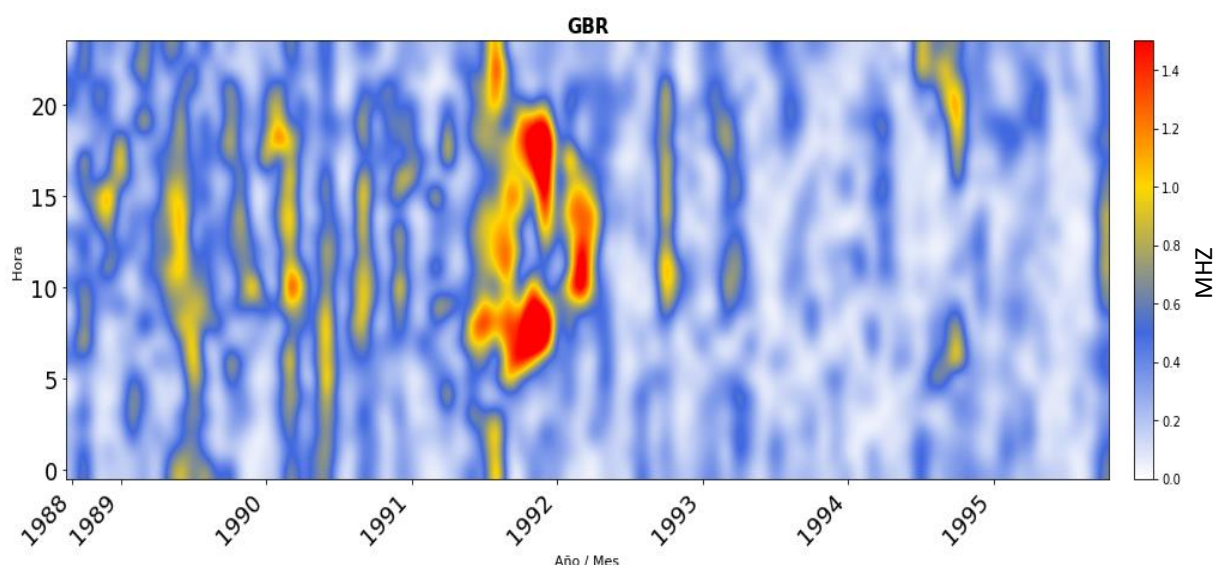


Figura 5.14 Valores de error absoluto para cada año-mes-hora del periodo interpolado utilizando el método GBR

Los valores promedios totales de los estadísticos resultan de nuevo similares a los obtenidos usando los modelos de regresión. En este caso estos valores son: 0.373 MHz para MAE, 6.98 % para MAPE y 0.961 para R^2 . Estos mismos valores se listan en la Tabla 7, más adelante, junto con los valores obtenidos usando los otros métodos de interpolación. Cabe destacar que, aunque levemente, estos valores reflejan una mejor performance que en el caso de RFR.

Valores y figuras similares a las anteriores se obtienen en el caso de la serie (2) interpolando los valores faltantes con GBR. Lo mismo ocurre con la serie (3) donde se generan 60% y 70% de huecos aleatorios.

En el caso de 90% y 97% de datos faltantes, los estadísticos se listan en la Tabla 5, donde nuevamente se nota una mejora en la performance de la interpolación con respecto a RFR.

Tabla 5 Valor de MAE, MAPE y R^2 para la serie de foF2 donde se generaron 90% y 97% de huecos que fueron luego interpolados con GBR.

	90% de huecos aleatorios	97% de huecos aleatorios
R^2	0.962	0.917
MAE (MHz)	0.306	0.437
MAPE (%)	6.15	8.87

5.4 Interpolación de foF2 utilizando Redes Neuronales

En este caso, como ya se mencionó se implementó una red neuronal artificial tipo perceptrón multicapa (MLP). Este enfoque de aprendizaje profundo es particularmente adecuado para modelar la compleja variabilidad ionosférica ya que permite capturar relaciones no lineales de alto orden entre múltiples proxies solares, geomagnéticos y variables temporales.

Los datos se separaron en dos conjuntos principales: uno conformado por los registros con valores conocidos de foF2 (utilizados para el entrenamiento) y otro con los registros faltantes (destinados a la predicción). El modelo fue entrenado utilizando las mismas variables predictoras que en los otros métodos de ML, es decir: F10.7, MgII, Rz, Ap, F30, las funciones seno y coseno de día y meses.

La arquitectura de la red neuronal consistió en una red secuencial con cuatro capas: una capa de entrada con 64 neuronas y activación ReLU, seguida de capas ocultas de 32 y 16 neuronas también con activación ReLU. y una capa de salida lineal con una única neurona para la predicción de foF2. El modelo fue compilado con el optimizador Adam y la función de pérdida mean squared error (MSE). El entrenamiento se realizó durante 50 épocas con un tamaño de batch de 64. También se utilizó como función de activación tanh. Los resultados fueron idénticos que con ReLU.

En la Figura 5.15 se muestran graficas de calor del MAE y MAPE para cada hora y mes, como en los casos anteriores. La Figura 5.16 muestra los valores del error absoluto para cada hora y mes/año. Se observan valores similares al caso de GBR, y a su vez, una mejora con respecto al método de RFR (ver Figuras 5.11 y 5.12 en el caso de RFR, y 5.13 y 5.14 en el caso de GBR).

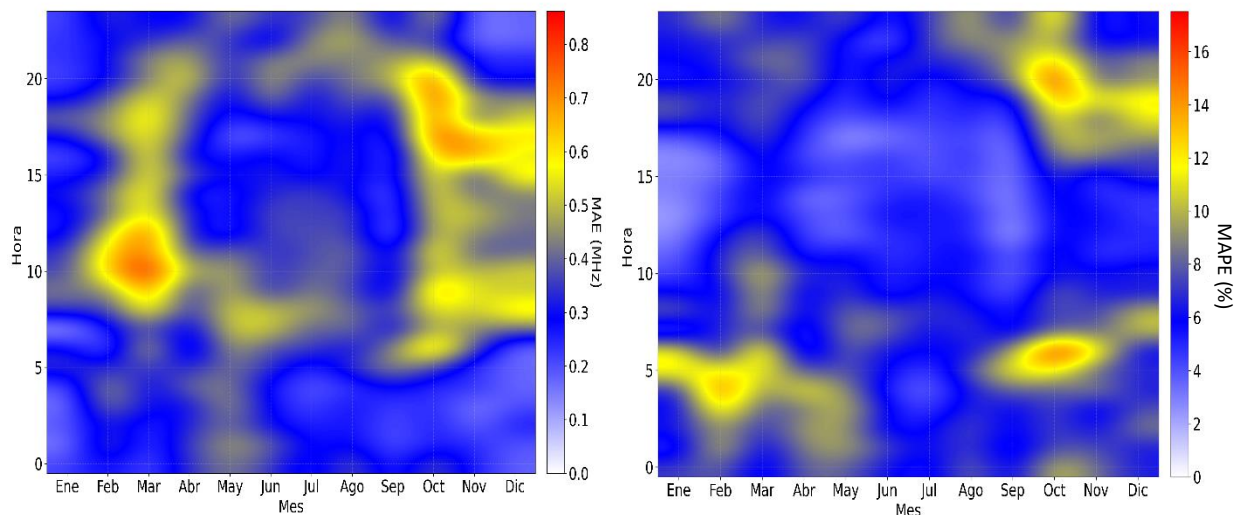


Figura 5.15 MAE y MAPE por hora y mes correspondiente a la técnica de redes Neuronales para la interpolación en el caso de la serie (1).

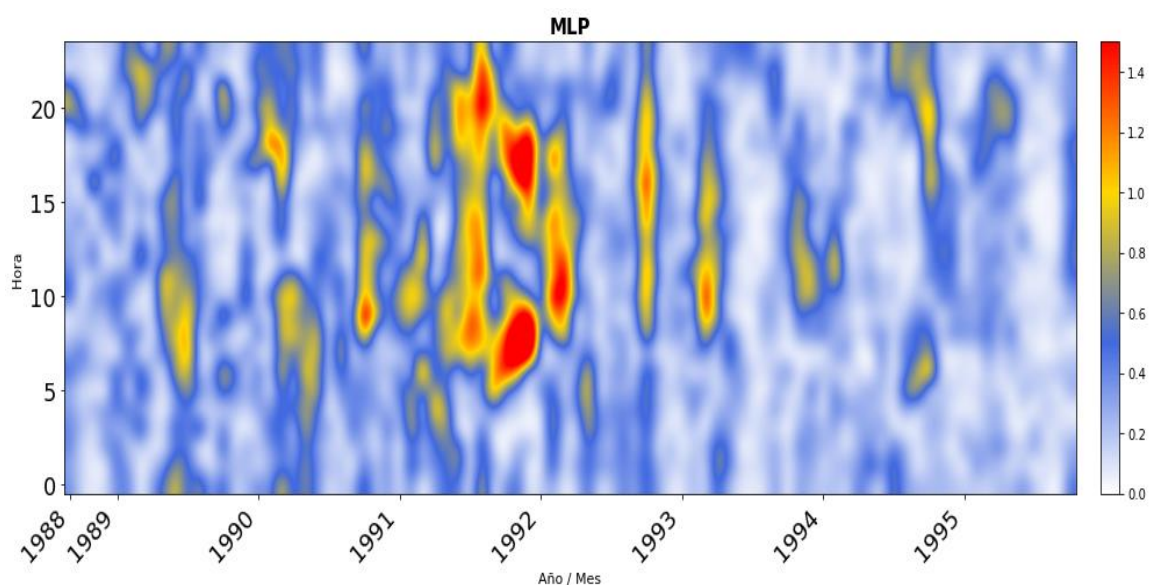


Figura 5.16 Valores de error absoluto para cada año-mes-hora del periodo interpolado utilizando el método de Redes Neuronales.

Los valores promedios totales de los estadísticos resultan, al igual que en los casos de RFR y GBR, similares a los obtenidos usando los modelos de regresión. En este caso de Redes Neuronales, los valores resultan: 0.358 MHz para MAE, 6.50 % para MAPE y 0.965 para R^2 . Estos mismos valores se listan en la Tabla 7 más adelante, junto con los valores obtenidos usando los otros métodos de interpolación.

En el caso de 90% y 97% de datos faltantes, los estadísticos se listan en la Tabla 6 donde nuevamente se nota una mejora en la performance de la interpolación con respecto a RFR, pero no respecto de GBR.

Tabla 6 Valor de MAE, MAPE y R^2 para la serie de foF2 donde se generaron 90% y 97% de huecos que fueron luego interpolados con Redes Neuronales.

	90% de huecos aleatorios	97% de huecos aleatorios
R^2	0.960	0.926
MAE (MHz)	0.327	0.448
MAPE (%)	6.78	9.49

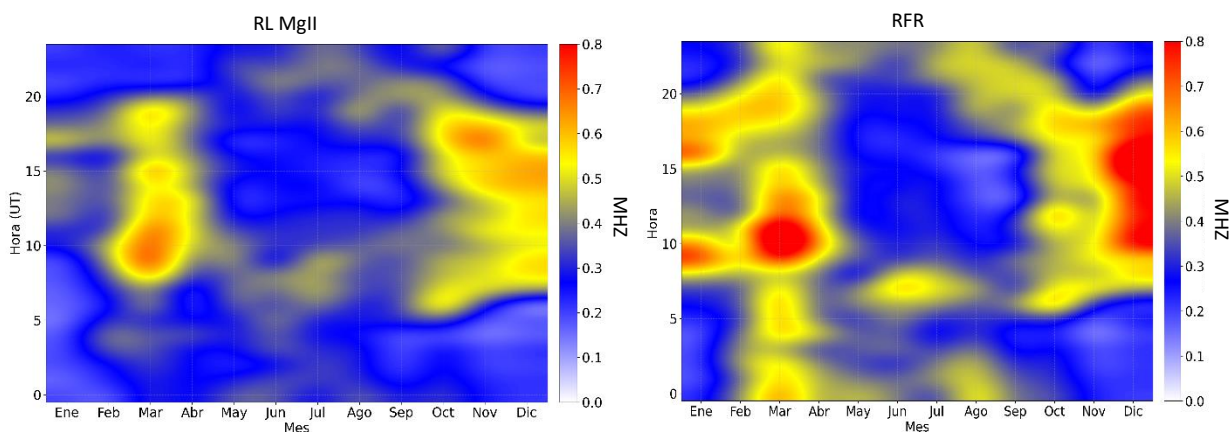
5.5 Análisis comparativo del desempeño de cada método de interpolación

En la Tabla 7 se listan los estadísticos comparativos para todos los modelos de interpolación utilizados en este trabajo, en el caso de la serie (1). Según los valores de error y R^2 , el método de regresión lineal es el de mejor performance. Y entre los métodos más complejos de ML, el de Redes Neuronales es el que arroja los valores óptimos de error y R^2 . La ventaja de estos últimos estaría en que no hay necesidad de separar la serie de foF2 en 288 series para tener en cuenta la variación diurna y estacional que no está presente en ninguno de los proxies que se consideran como predictores.

Tabla 7. Valor de MAE, MAPE y R^2 para la serie (1) obtenido con cada uno de los métodos de interpolación considerados.

Modelo	RL Mg2	RFR	GBR	Redes Neuronales
R^2	0.966	0.951	0.961	0.965
MAE (MHz)	0.345	0.409	0.373	0.358
MAPE (%)	6.04	7.44	6.98	6.50

En la Figura 5.17 se agrupan las gráficas de MAE considerando separadamente horas y meses. Se utilizó una escala común a todos los paneles, de manera que se observa claramente que el modelo de desempeño más pobre es RFR. En todos se observa que las horas y meses de peor performance en la interpolación de todos los métodos corresponde a los equinoccios de invierno y hora entre 10 y 18 LT que es cuando los valores de foF2 son mayores. En el caso de MAPE, estos valores se corren a horas de la noche.



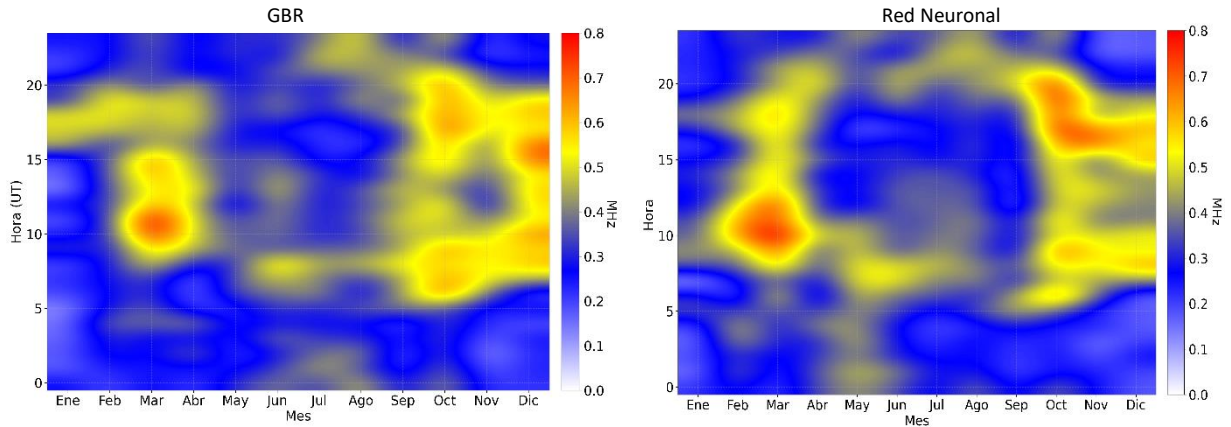


Figura 5.17 Valores de MAE por hora y mes para la serie (1) y para los cuatro métodos de interpolación considerados.

La Figura 5.18 muestra los valores de error en función del mes/año y hora, es decir sin ningún tipo de promedio. Nuevamente, y como es de esperarse, lo mayores errores ocurren para las épocas donde foF2 alcanza los valores más grandes, es decir alrededor del mediodía y tarde, y para los años de actividad solar máxima, es decir alrededor del año 1991. En el caso de RFR se observan zonas rojas (es decir con error del orden de 1MHz) en el período previo al máximo de actividad solar.

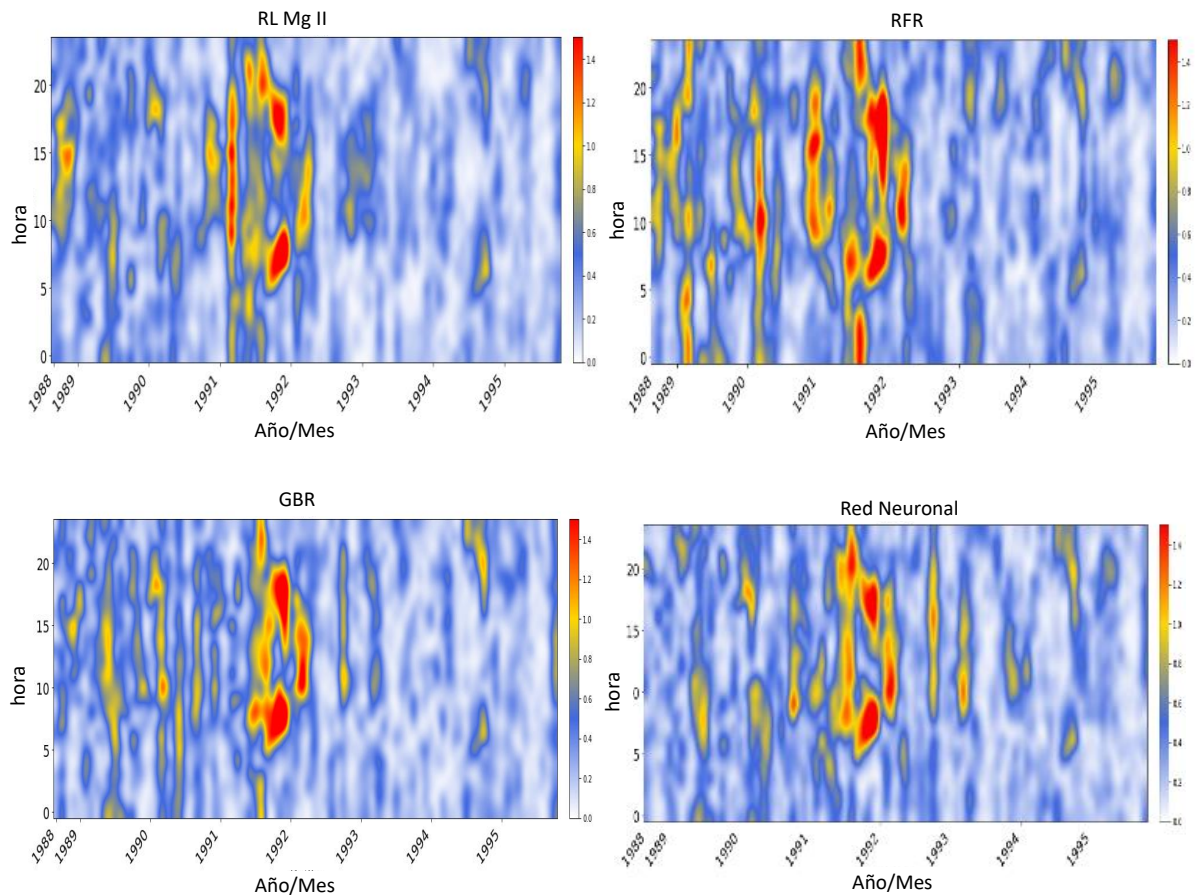


Figura 5.18 Error absoluto (MHz) en función del mes/año y hora de la serie (1) para los cuatro métodos de interpolación considerados, en el período de interpolación.

La Tabla 8 muestra los valores de MAE, MAPE y R^2 obtenidos al aplicar los métodos RFR, GBR, Redes Neuronales y RL sobre la serie (1), pero restringiendo el promedio a los intervalos horarios de 0 a 4 LT y de 0 a 1 LT. Esta segmentación permite evaluar el comportamiento de los modelos en los periodos donde la correlación entre foF2 y los proxies solares es menor a 0.9. En estos casos se observa que la mejor performance la tienen los modelos más complejos de ML, superando el desempeño de la regresión lineal.

Tabla 8. Valor de MAE, MAPE y R^2 para la serie (1) obtenido con los métodos RFR, GBR, Redes Neuronales e RL MgII, promediando para horas de 0 a 4 LT y 0 a 1 LT.

Periodo 1988-1995	Modelo de interpolación	RFR	GBR	Redes Neuronales	RL Mg II
Hora 0 a 4	R^2	0.866	0.919	0.911	0.892
	MAE (MHz)	0.342	0.275	0.282	0.312
	MAPE (%)	9.12	7.50	7.57	8.24
Hora 0 a 1	R^2	0.861	0.923	0.915	0.898
	MAE (MHz)	0.376	0.282	0.293	0.330
	MAPE (%)	9.24	7.25	7.43	8.04

En la Tabla 9 se agrupan los valores de los estadísticos comparativos para el caso de 90% y 97% de huecos generados de forma aleatoria. En este caso solo se consideran los modelos de ML más complejos: RFR, GBR y Redes Neuronales.

Tabla 9. Valor de MAE, MAPE y R^2 para la serie (3) obtenido con los métodos RFR, GBR, Redes Neuronales e IL MgII, con 60% y 70% de huecos generados de manera aleatoria.

	Modelo de interpolación	RFR	GBR	Redes Neuronales	RL Mg II
60% de huecos aleatorios	R^2	0.973	0.982	0.975	0.956
	MAE (MHz)	0.255	0.214	0.265	0.342
	MAPE (%)	5.18	4.41	5.45	6.78
70% de huecos aleatorios	R^2	0.965	0.982	0.970	0.950
	MAE (MHz)	0.283	0.214	0.281	0.358
	MAPE (%)	5.72	4.41	5.67	7.10

Tabla 10. Valor de MAE, MAPE y R^2 para la serie (3) obtenido con los métodos RFR, GBR y Redes Neuronales, con 90% y 97% de huecos generados de manera aleatoria.

	Modelo de interpolación	RFR	GBR	Redes Neuronales
--	-------------------------	-----	-----	------------------

90% de huecos aleatorios	R²	0.938	0.962	0.960
	MAE (MHz)	0.386	0.306	0.327
	MAPE (%)	7.76	6.15	6.78
97% de huecos aleatorios	R²	0.904	0.917	0.926
	MAE (MHz)	0.486	0.437	0.448
	MAPE (%)	10.00	8.87	9.49

Capítulo 6. Discusión y conclusiones

Los métodos de interpolación de datos faltantes en un parámetro ionosférico con distintos niveles de complejidad explorados en esta Tesis junto con su análisis comparativo permiten valorar las ventajas y limitaciones de los enfoques clásicos y los basados en aprendizaje automático. En este último capítulo se analizan los resultados obtenidos y su aporte al estudio de la ionosfera. Se presentan las discusiones y conclusiones generales y finalmente se plantean las líneas futuras de investigación que quedan abiertas a partir de este trabajo.

6.1 Método de interpolación más conveniente en el caso del parámetro ionosférico foF2: simpleza vs. precisión

En esta sección se realiza un análisis comparativo del desempeño de los distintos métodos de interpolación evaluados: Regresión Lineal (RL) con el índice Mg II como variable predictora, Random Forest Regression (RFR), Gradient Boosting Regression (GBR) y Redes Neuronales. La comparación se fundamenta en las métricas de error MAE, MAPE y el coeficiente de determinación R^2 , obtenidas al aplicar los modelos sobre series de datos con diferentes configuraciones de huecos, representando escenarios de diversa complejidad y naturaleza.

La Tabla 7 presenta los resultados para la serie con un hueco continuo de 7 años (~10% de los datos). En este escenario, todos los métodos demostraron un rendimiento notablemente alto, con valores de R^2 superiores a 0.95. La RLMg II mostró el mejor desempeño general, alcanzando el R^2 más alto (0.966) y los menores errores (MAE = 0.345 MHz, MAPE = 6.04%). El modelo GBR también exhibió un rendimiento excelente (R^2 = 0.961, MAPE = 6.98%), superando al RFR y a las Redes Neuronales, que obtuvieron resultados ligeramente inferiores pero aún muy aceptables.

Sin embargo, al analizar subperiodos críticos como las horas donde la correlación de foF2 con los proxies solares cae significativamente (0 a 4 y, más específicamente, 0 a 1 hora) en la Tabla 8, se observa un cambio en el panorama. En estos intervalos, el GBR emerge como el método más robusto, superando al resto en todas las métricas. Para el periodo de 0 a 1 hora, el GBR obtuvo un R^2 de 0.923 y un MAPE de 7.25%, superando al RL Mg II (R^2 =0.898, MAPE=8.04%) y a las Redes Neuronales (R^2 =0.915, MAPE=7.43%). Este comportamiento sugiere que, aunque la RL Mg II es altamente efectiva en condiciones globales, los métodos basados en árboles de decisión potenciados (GBR) poseen una mayor capacidad para modelar las no linealidades sutiles y las interacciones entre variables que se manifiestan en los regímenes de baja actividad, donde la dependencia lineal con el índice Mg II podría ser menos pronunciada.

El análisis de la Serie 3, donde los huecos son de naturaleza aleatoria, proporciona la evaluación más rigurosa de la capacidad de generalización de los modelos, ya que este escenario simula pérdidas de datos impredecibles y no correlacionadas temporalmente. Los resultados de la Tabla 9 son reveladores. En el escenario del 60% de huecos aleatorios, todos los métodos mostraron un desempeño excelente, con R^2 superiores a 0.95. Sin embargo, el GBR se destaca claramente, logrando el mejor R^2 (0.982), el MAE más bajo (0.214 MHz) y el MAPE más bajo (4.41%) para ambos porcentajes de huecos (60% y 70%). Este rendimiento es superior al de la RL Mg II ($R^2 \approx 0.95$, MAPE $\approx 7\%$), al RFR y a las Redes Neuronales. La capacidad del GBR

para construir modelos secuenciales que corrigen los errores de sus predecesores lo convierte en una herramienta eficiente para interpolar datos faltantes cuando la información disponible está dispersa, pero en una proporción aún manejable. La RL Mg II, aunque competente, muestra sus limitaciones al necesitar una relación lineal más estricta, evidenciada por su mayor MAPE en estos escenarios.

La Tabla 10 presenta un desafío mayor: escenarios con el 90% y 97% de los datos faltantes de forma aleatoria. Aquí se observa una degradación esperada en el rendimiento de todos los modelos, pero con matices importantes. Para un 90% de huecos, el GBR y las Redes Neuronales obtienen resultados prácticamente idénticos y muy buenos ($R^2 \approx 0.96$, MAPE ≈ 6.1 -6.8%), superando al RFR. Sorprendentemente, cuando la proporción de datos faltantes alcanza el 97%, las Redes Neuronales superan ligeramente al GBR, logrando el mejor R^2 (0.926) y un MAPE de 9.49%, frente al 8.87% del GBR. Este resultado sugiere que, en condiciones de información extremadamente escasa, la arquitectura de las Redes Neuronales, con su capacidad para aprender representaciones jerárquicas y complejas de los datos, puede generalizar de manera más efectiva que los métodos de ensamble, los cuales pueden volverse más susceptibles al sobreajuste cuando la cantidad de datos de entrenamiento es críticamente baja.

El análisis integral de los resultados permite extraer conclusiones sobre la idoneidad de cada método según el escenario. En primer lugar, la RL Mg II es un método simple, eficiente y altamente competitivo cuando la relación física subyacente es lineal y los huecos no son excesivamente complejos (Serie 1). Su desempeño se degrada en los regímenes de baja actividad y cuando la cantidad de datos faltantes es alta y aleatoria (Serie 3), donde su incapacidad para modelar no linealidades se vuelve una limitación. En segundo lugar, el GBR se posiciona como el método más versátil y robusto en la mayoría de los escenarios evaluados. Ofrece un rendimiento excelente en huecos continuos (especialmente en horas de baja actividad), y es el indiscutible ganador en los escenarios de huecos aleatorios moderados (60-70%), mostrando el menor error y la mayor precisión. Su capacidad para capturar interacciones complejas y su enfoque secuencial de corrección de errores le otorgan una ventaja significativa. En tercer lugar, las Redes Neuronales muestran un rendimiento muy bueno y estable en todos los casos, aunque sin superar consistentemente al GBR, excepto en el escenario más extremo (97% de huecos aleatorios). Su capacidad para manejar grandes volúmenes de datos y aprender estructuras no lineales profundas las convierte en una opción muy potente, especialmente cuando la cantidad de información disponible es críticamente baja, aunque su implementación y ajuste de hiperparámetros pueden ser más demandantes. Finalmente, el RFR si bien ofrece resultados aceptables, generalmente es superado por el GBR y las Redes Neuronales en todas las métricas, lo que sugiere que, para este conjunto de datos, los métodos de boosting y las arquitecturas neuronales poseen una mayor capacidad de modelado.

En conclusión, para la interpolación de la serie temporal estudiada, el método GBR emerge como la opción más recomendable debido a su equilibrio entre precisión y robustez en una amplia gama de escenarios de datos faltantes. La Regresión Lineal con Mg II sigue siendo una excelente opción de referencia por su simplicidad y efectividad en condiciones ideales, mientras que las Redes Neuronales se perfilan como la alternativa más adecuada para los casos de pérdida de datos masiva y aleatoria.

6.2 Aportes de los resultados de esta Tesis a la problemática actual de las series de datos con huecos y a la necesidad de series completas

Las series de tiempo de mediciones constituyen una herramienta fundamental en numerosas áreas de investigación, incluyendo ciencias ambientales, climatología, economía, y física. Sin embargo, la presencia de datos faltantes es un problema recurrente debido a fallas instrumentales, interrupciones operativas, errores de medición o dificultades en la adquisición de observaciones. La existencia de huecos limita la aplicación de numerosos métodos estadísticos y aumenta la incertidumbre asociada a los resultados obtenidos, pudiendo afectar significativamente las conclusiones derivadas del análisis de los datos (Schneider, 2001; Little & Rubin, 2002).

En estudios basados en series de largo plazo, la disponibilidad de registros completos adquiere una relevancia aún mayor. Trabajos recientes han mostrado que la ausencia de observaciones puede introducir sesgos en la estimación de valores medios y tendencias, afectando incluso la interpretación de señales de largo plazo cuando la cantidad de datos faltantes supera ciertos umbrales (Elias et al., 2026). Por esta razón, el desarrollo de metodologías robustas para la reconstrucción de datos faltantes constituye actualmente una línea de investigación de interés en diversas disciplinas científicas.

Los resultados obtenidos en esta Tesis muestran que las técnicas de aprendizaje automático permiten estimar con elevada precisión valores ausentes en series temporales de foF2, incluso bajo escenarios de pérdida sustancial de información. En este sentido, el presente trabajo aporta evidencia adicional acerca del potencial de estas herramientas para mejorar la continuidad y calidad de las bases de datos experimentales, favoreciendo la utilización de registros históricos en estudios estadísticos, climatológicos y de modelado físico.

6.3 Comparación con otros trabajos que utilizan técnicas de ML para predecir valores de parámetros ionosféricos

Durante los últimos años se ha observado un crecimiento sostenido en la aplicación de técnicas de ML al estudio de la ionosfera. Diversos trabajos han demostrado que métodos como Random Forest, Gradient Boosting, Support Vector Machines y Redes Neuronales son capaces de reproducir con alta precisión la variabilidad temporal de parámetros ionosféricos, incluyendo foF2 (Pan et al., 2021; Muzaffer & Senturk, 2022; Zhou et al., 2024; Zossi et al., 2025).

No obstante, los resultados también muestran que modelos relativamente simples pueden alcanzar niveles de precisión satisfactorios cuando existe una adecuada selección de variables predictoras como lo demuestra este trabajo utilizando Regresión Lineal. Esto sugiere que la complejidad creciente de los modelos no siempre implica mejoras proporcionales en el desempeño, particularmente cuando se busca una solución computacionalmente eficiente y físicamente interpretable. A diferencia de muchos trabajos previos enfocados en la predicción de valores futuros, esta investigación se concentra en la reconstrucción de datos faltantes dentro de una serie observacional, ampliando así el campo de aplicación de las técnicas de aprendizaje automático en Física Espacial.

Algo que debemos destacar es que la mayoría de los trabajos de otros autores analizan datos ionosféricos con mayor frecuencia temporal, donde los proxies solares ya no presentan correlaciones tan altas como en el caso de la variabilidad interanual, y en donde se esperaría que el método simple de regresión lineal pierda la superioridad en performance frente a los modelos más complejos de ML que logran modelar asociaciones no-lineales.

Los resultados de esta Tesis podrían ser comparables con los resultados de Zossi et al. (2025), aunque en este trabajo se utiliza la metodología de RFR y GBR para completar datos faltantes reales (es decir que no se cuenta con los datos observados para verificar). Pero se trata del único trabajo publicado que analiza las técnicas usadas en esta Tesis para completar datos de medianas mensuales. En Zossi et al. (2025) se concluye que los modelos de GBR y RFR son superiores que la Regresión Lineal.

Se insiste con el hecho de que en este trabajo los modelos de RFR, GBR y Redes Neuronales se mantuvieron en sus configuraciones más simples sin utilizar múltiples optimizadores que Python u otro lenguaje de programación pone a disposición.

6.4 Líneas futuras de investigación

Los resultados obtenidos en este trabajo abren diversas posibilidades de extensión y profundización que permitirían consolidar y generalizar la metodología desarrollada.

En primer lugar, resultaría de interés extender el análisis a un conjunto más amplio de estaciones ionosféricas, con el objetivo de evaluar si la elección del método de interpolación más adecuado depende de la región geográfica. La latitud de la estación constituye un factor potencialmente determinante: en estaciones de alta latitud, la actividad geomagnética, representada en este trabajo por el índice Ap, ejerce una influencia más pronunciada sobre la ionosfera debido a la mayor penetración de partículas energéticas y a la presencia de corrientes aurales, mientras que en estaciones de baja latitud dicha influencia puede resultar secundaria frente al forzante de radiación EUV solar. Un estudio comparativo sistemático permitiría establecer criterios objetivos para la selección del método en función de la posición geográfica de la estación, y en particular determinar en qué regímenes de latitud la inclusión del índice Ap aporta una mejora significativa al interpolado.

En segundo lugar, el presente análisis se centró en datos con resolución mensual para cada hora del día. Extender el estudio a resolución diaria plantea desafíos adicionales, ya que a esa escala temporal la actividad geomagnética adquiere mayor relevancia como variable explicativa, al tiempo que la variabilidad ionosférica resulta menos predecible a partir de índices de actividad solar. Esto podría requerir una revisión tanto de las variables de entrada consideradas como de la arquitectura de los modelos empleados, otorgando mayor peso relativo a índices geomagnéticos en detrimento de los proxies solares utilizados aquí.

En tercer lugar, una línea de trabajo de particular relevancia científica consiste en evaluar el impacto del método de relleno de huecos sobre el cálculo de otras variables derivadas de la serie de foF2. Un ejemplo concreto es la estimación de tendencias de largo plazo mediante técnicas estadísticas que requieren series temporales completas y continuas. La calidad del

interpolado puede introducir sesgos sistemáticos o modificar artificialmente la varianza de la serie, afectando potencialmente las conclusiones sobre tendencias atribuibles a cambios en la composición atmosférica o en la actividad solar a escalas decadales. Cuantificar este efecto constituye un paso necesario hacia una caracterización más rigurosa de la ionosfera a largo plazo.

Puede resultar de interés también analizar métodos de interpolación en función de la misma serie de datos, es decir sin utilizar parámetros externos como lo son los proxies solares. Existen métodos basados en la Transformada de Fourier, SSA (Singular Spectrum Analysis), ARIMA (Autoregressive Integrated Moving Average) entre otros, que permiten predecir una variable en función del comportamiento los datos disponibles de la misma.

Finalmente, en este trabajo los modelos RFR, GBR y MLP fueron implementados en sus configuraciones más elementales, priorizando deliberadamente la transparencia matemática de cada método por sobre la optimización del rendimiento. Esta decisión permitió mantener un control riguroso sobre el funcionamiento interno de los algoritmos y su comprensión a nivel fundamental. Como extensión natural, una línea futura consiste en explorar de manera sistemática el conjunto de optimizadores, estrategias de regularización y técnicas de búsqueda de hiperparámetros disponibles en los entornos de programación científica actuales con el único objetivo de llevar el rendimiento de cada modelo a su límite, sin la restricción de mantener plena conciencia de cada componente involucrado. Este enfoque permitiría establecer un techo de rendimiento empírico para cada método y evaluar en qué medida la simplicidad adoptada en este trabajo implica un costo en términos de precisión del interpolado.

Referencias

- Balan, N., Bailey, G.J., Jenkins, B., Rao, P.B., Moffet, R.J. (1994). Variations of ionospheric ionization and related solar fluxes during an intense solar cycle, *J. Geophys. Res.*, 99, 2243–2253. [10.1029/93JA02099](https://doi.org/10.1029/93JA02099)
- Bilitza, D. Pezzopane, M., Truhlik, V., Altadill, D., Reinisch, B. W., & Pignalberi, A. (2022). The International Reference Ionosphere model: A review and description of an ionospheric benchmark. *Reviews of Geophysics*, 60, e2022RG000792. <https://doi.org/10.1029/2022RG000792>
- Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Camporeale, E., Wing, S., Johnson, J.R. (2018). *Machine Learning Techniques for Space Weather*, Elsevier, Amsterdam.
- Camporeale, E. (2019). The challenge of machine learning in Space Weather: Nowcasting and forecasting. *Space Weather*, 17, 1166–1207.
- Chapman, S. (1931). The absorption and dissociative or ionizing effect of monochromatic radiation in an atmosphere on a rotating earth. *Proceedings of the Physical Society*, 43, 26. DOI 10.1088/0959-5309/43/1/305
- Chen, Y. Liu, L., Le, H., Zhang, H. Zhang, R. (2022). Responding trends of ionospheric F2-layer to weaker geomagnetic activities, *J. Space Weather Space Clim.*, 12, 6.
- Clette, F., Svalgaard, L., Vaquero, J.M., Cliver, E. W. (2014). Revisiting the Sunspot Number. A 400-Year Perspective on the Solar Cycle, *Space Science Reviews*, 186, 35–103. DOI: 10.1007/s11214-014-0074-2
- Damboldt, T., Suessmann, P. (2012). Consolidated Database of Worldwide Measured Monthly Medians of Ionospheric Characteristics foF2 and M(3000)F2. INAG (Ionosonde Network Advisory Group) Bulletin 73 https://www.ursi.org/files/CommissionWebsites/INAG/web-73/2012/damboldt_consolidated_database.pdf
- Dudok de Wit, T., S. Bruinsma, K. Shibasaki (2014). Synoptic radio observations as proxies for upper atmosphere modelling. *J. Space Weather Space Clim.*, 4, A06. DOI: 10.1051/swsc/2014003.
- Elias, A.G. (2014). Filtering ionosphere parameters to detect trends linked to anthropogenic effects. *Earth. Planet Sp.*, 66, 113. <https://doi.org/10.1186/1880-5981-66-113>
- Elias, A. G., Martinis, C. R., de Haro Barbas, B. F., Medina, F. D., Zossi, B. S., Fagre, M., and Duran, T. (2023). Comparative analysis of extreme ultraviolet solar radiation proxies during minimum activity levels. *Earth Planet. Phys.* 7, 540–547. <https://doi.org/10.26464/epp2023050>.
- Elias, A.G., Zossi, B.S., Silvero, C.I., Zamora, J., Duran, T. & Abaca, F. (2026). Data gaps and trend bias relevant to space climate studies. *Discover Space*, 130, 1.

- Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, 29, 1189-1232. <https://doi.org/10.1007/s11038-026-09580-8>
- Gueymard, C.A. (2014). A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects, *Renewable and Sustainable Energy Reviews*, 39, pp. 1024–1034. <https://doi.org/10.1016/j.rser.2014.07.117>
- Hargreaves, J.K. (1992). *The solar-terrestrial environment*, Cambridge University Press, Cambridge, UK.
- Heath, D. F., and B. M. Schlesinger (1986), The Mg 280-nm doublet as a monitor of changes in solar ultraviolet irradiance, *J. Geophys. Res.*, 91(D8), 8672–8682, doi:10.1029/JD091iD08p08672
- Kowalek, P., Loch-Olszewska, H., Szwabiński, J. (2019). Classification of diffusion modes in single-particle tracking data: feature-based versus deep-learning approach. *Phys. Rev. E*, 100, 032410. <https://doi.org/10.1103/PhysRevE.100.032410>
- Lastovicka, J., Jelinek, S. (2019). Problems in calculating long-term trends in the upper atmosphere, *J. Atmos. Sol.-Terr. Phy.*, 189, 80–86- <https://doi.org/10.1016/j.jastp.2019.04.011>
- Lastovicka, J., Buresova, D. (2023). Relationship between foF2 and solar activity proxies. *Space Weather*, 21, e2022SW003359. <https://doi.org/10.1029/2023SW003359>
- Lastovicka, J. (2024). Dependence of long-term trends in foF2 at middle latitudes on different solar activity proxies. *Adv. Space Res.*, 73, 685-689. <https://doi.org/10.1016/J.ASR.2023.09.047>.
- Little, R.J.A. and D.B. Rubin (2002). *Statistical analysis with missing data*. 2nd edn. Hoboken, John Wiley & Sons, Inc.
- Liu, L., W. Wan, and B. Ning (2004), Statistical modeling of ionospheric foF2 over Wuhan, *Radio Sci.*, 39, RS2013, doi:10.1029/2003RS003005
- Liu L, Wan W, Ning B, Pirog OM, Kurkin VI (2006). Solar activity variations of the ionospheric peak electron density. *J. Geophys. Res.*, 111, A08304, doi:10.1029/2006JA011598 111
- Matzka, J., Stolle, C., Yamazaki, Y., Bronkalla, O., & Morschhauser, A. (2021). The geomagnetic Kp index and derived indices of geomagnetic activity. *Space Weather*, 19, e2020SW002641. <https://doi.org/10.1029/2020SW002641>
- Medina, F. D., Zossi, B. S., Bossolasco, A., and Elias, A.G. (2023). Performance of CHIRPS dataset for monthly and annual rainfall-indices in Northern Argentina, *Atmospheric Research*, 283, 106545. <https://doi.org/10.1016/j.atmosres.2022.106545>
- Mishra, A. (2020). *Machine Learning for IOS Developers*, Wiley, Oxford, UK.
- Muzaffer, C.I., Senturk, E. (2022). Machine learning regression models for prediction of multiple ionospheric parameters, *Adv. Space Res.*, 69, 1319–1334.
- Pan, Y., M. Jin, S. Zhang, and Y. Deng (2021), TEC Map Completion Through a Deep Learning Model: SNP-GAN, *Space Weather*, 19(11), e2021SW002810.

- Picone, J. M., A. E. Hedin, D. P. Drob, A. C. Aikin (2002). NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues, *J. Geophys. Res.*, 107, 1468. doi:10.1029/2002JA009430
- Piggott, W.R., Rawer, K. (1978). U.R.S.I. handbook of ionogram interpretation and reduction. 2nd edition. Washington, D.C., U.S. Department of Commerce.
- Radicella, S.M. (2023). New Ways to Modelling and Predicting Ionosphere Variables. *Atmosphere*, 14, 1788. <https://doi.org/10.3390/atmos14121788>
- Reinisch, B. W., R. R. Gamache, J. S. Tang, D. F. Kitrosser (1983), Automatic real time ionogram scaler with true height analysis-ARTIST, Tech. Rep., DTIC Document. [Available at www.dtic.mil/dtic/tr/fulltext/u2/a135174.pdf.]
- Reinisch, B. W., I. A. Galkin, G. M. Khmyrov, A. V. Kozlov, K. Bibl, I. A. Lisysyan, G. P. Cheney, X. Huang, D. F. Kitrosser, V. V. Paznukhov, Y. Luo, W. Jones, S. Stelmash, R. Hamel, J. Grochmal (2009), New Digisonde for research and monitoring applications, *Radio Sci.*, 44, RS0A24, doi:10.1029/2008RS004115.
- Reinisch, B. W., I. A. Galkin (2011). Global ionospheric radio observatory (GIRO), *Earth, Planets, and Space*, 63, 377-381. doi:10.5047/eps.2011.03.001
- Rishbeth, H., Garriott, O.K. (1969). Introduction to ionospheric physics, Academic Press, New York, USA.
- Schneider T. (2001). Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.*, 14, 853–71. [https://doi.org/10.1175/1520-0442\(2001\)014<0853:AOICDE>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2)
- Schunk, R.W., Nagy, A.F. (2009). Ionospheres. Physics, Plasma Physics, and Chemistry, Cambridge University Press, Cambridge, UK.
- Snow, M., Weber, M., Machol, J., Viereck, R., and Richard, E. (2014). Comparison of Magnesium II core-to-wing ratio observations during solar minimum 23/24. *J. Space Weather Space Clim.*, 4, A04. <https://doi.org/10.1051/swsc/2014001>
- Tan Jun Rios, M. G., Borries, C., Liu, H., Mielich, J. (2025). Long-term changes in the dependence of NmF2 on solar flux at Juliusruh, *Ann. Geophys.*, 43, 73–89, <https://doi.org/10.5194/angeo-43-73-2025>
- Tapping, K. F. (2013), The 10.7 cm solar radio flux (F10.7), *Space Weather*, 11, 394–406, doi:10.1002/swe.20064
- Zhou, Y., Liu, J., Li, S., Li, Q. (2024). Ionospheric TEC prediction based on ensemble learning models. *Space Weather*, 22, e2023SW003790. <https://doi.org/10.1029/2023SW003790>
- Zolessi, B., Cander, L.R. (2014). Ionospheric Prediction and Forecasting, Springer-Verlag, Berlin, Germany.
- Zossi, B.S., Medina, F.D., Duran, T., Fagre, M., de Haro Barbas, B.F., Elias, A.G. (2025). Filling ionospheric monthly medians missing data: a machine learning approach, *Proc. R. Soc. A*.48120240566. <http://doi.org/10.1098/rspa.2024.0566>