UNIVERSIDAD AUTÓNOMA DE NAYARIT
POSGRADO EN CIENCIAS BIOLÓGICO AGROPECUARIAS

PROPUESTA METODOLÓGICA PARA EL MONITOREO DE LA CONCENTRACIÓN DE CLOROFILA-A EN AGUAS CONTINENTALES POR MEDIO DE SENSORES REMOTOS

LIZETTE ZAREH CORTES MACÍAS

Tesis presentada como requisito parcial para la obtención del grado de:
Maestría en Ciencias Biológico Agropecuarias en el Área de Ciencias Ambientales

Xalisco, Nayarit. Julio, 2023.

UNIVERSIDAD AUTÓNOMA DE NAYARIT
POSGRADO EN CIENCIAS BIOLÓGICO AGROPECUARIAS

PROPUESTA METODOLÓGICA PARA EL MONITOREO DE LA CONCENTRACIÓN DE CLOROFILA-A EN AGUAS CONTINENTALES POR MEDIO DE SENSORES REMOTOS

LIZETTE ZAREH CORTES MACÍAS

Tesis presentada como requisito parcial para la obtención del grado de:
Maestría en Ciencias en el Área de Ciencias Ambientales

Comité tutorial
Director: Dr. Juan Pablo Rivera Caicedo
Co-director: Dr. Jushiro C. A. Cepeda Morales
Tutores: Dr. Oscar Ubisha Hernández Almeida; Dr. Ricardo García Morales

Xalisco, Nayarit. Julio, 2023.

Xalisco, nayarit. A 29 de junio del 2023

DR. JUAN DIEGO GARCÍA PAREDES
COORDINADOR DE POSGRADO EN
CIENCIAS BIOLÓGICO AGROPECUARIAS (CBAP)
UNIVERSIDAD AUTÓNOMA DE NAYARIT
P R E S E N T E

Lo suscritos integrantes del Comité Tutorial del alumno C. **Biol. Lizette Zareh Cortes Macías**, después de revisar minuciosamente su trabajo de tesis titulado **"Propuesta metodológica para el monitoreo de la concentración de clorofila-*a* en aguas continentales por medio de sensores remotos"**, hemos determinado que la tesis cumple con los criterios de calidad y originalidad suficientes, por lo cual puede ser presentada por el alumno para aspirar al grado de Maestría en Ciencias con la opción terminal en Ciencias Ambientales.

Damos nuestra **APROBACIÓN** en cumplimiento a los lineamientos del reglamento de estudios del Posgrado en CBAP de la UAN vigentes con la finalidad de que el alumno continúe los trámites correspondientes para la obtención del grado.

ATENTAMENTE

DR. JUAN PABLO RIVERA CAICEDO
DIRECTOR

DR. JUSHIRO C.A. CEPEDA MORALES
CO-DIRECOR

DR. OSCAR U. HERNÁNDEZ ALMEIDA
ASESOR

DR. RICARDO GARCÍA MORALES
ASESOR

c.c.p. Interesado
c.c.p. Archivo

A la memoria de mis abuelitas: María y Alejandra, y mis abuelitos: Gregorio y Apolinar.

"(...) Jamás se dice adiós allá,
Jamás se dice adiós,
En el país de gozo y paz,
Jamás se dice adiós.
La voz de triste despedida
Jamás allí se oirá,
Mas la canción de paz y gozo
Por siempre durará. (...)"

En memoria de Chema.

Quería ser mensaje de cariño y libertad, quería ser mensaje.
Quería ser paisaje y dormir junto a la mar, quería ser paisaje.
Quería ser salvaje y no tenerse que atar, quería ser salvaje.
¿Cómo ser mensaje? Si no hay nadie a quien hablar,
¿Como ser mensaje?
¿Cómo ser paisaje? Junto a una planta de gas,
¿Como ser paisaje?
¿Y cómo ser salvaje? Entre mil cadenas,
¿Como ser salvaje? (...)

- El Último Ke Zierre, (1995)

Si me encierro, ven a verme; un vis a vis
Caí presa dentro de mí, dentro, muy dentro de mí.
Si me escapo, ve a buscarme cualquier día
Donde quede alguna flor, donde no haya policía.

- Extremoduro (1994)

Agradecimientos

A mi **mamá** y a mi **papá**, por tantas lecciones de resiliencia y fortaleza. Por brindarme un hogar, por ser mi ancla. **Los amo**.

Al doctor **Juan Pablo** y el doctor **Jushiro**, por estar conmigo en este proceso de crecimiento académico y personal. Por compartir su conocimiento conmigo, por darme el espacio y la confianza para desarrollarme profesionalmente. Gracias por todo su apoyo.

Al doctor **Ubisha** y al doctor **Ricardo** por formar parte de este proyecto, por sus aportaciones que ayudaron a enriquecer y aterrizar las ideas que buscaba plasmar.

A las personas que he conocido en **CENITT**, que ya es como una segunda casa para mí. A mis compañeros en **PERSEO**: Enrique y Alejandra que extraño tantooo. Gracias por su compañía y su apoyo, ¡qué gusto conocerles!

Gracias a la vida por mis amigas, que más que amigas son mis carnalas: **Liz** y **May**, malditas panks, las amo. Gracias por ser y por estar.

Gracias al Posgrado en Ciencias Biológico Agropecuarias y a CONACyT por facilitar la oportunidad de realizar un posgrado.

A las agencias espaciales que mantienen sus bases de datos de acceso público, DUREN.

Todo este tiempo pensé sobre qué iba a escribir en mis agradecimientos, creí que me haría falta espacio para enlistar todas las razones por las que estoy agradecida, sin embargo, ahora creo que las palabras sobran. Estoy agradecida por estar aquí, porqué la vida sí es chida. Agradezco estar rodeada de personas que me inspiran a ser mejor cada día. Y estoy muy feliz de cerrar este ciclo.

Índice general

Índice de figuras

Índice de tablas

Resumen

Los cuerpos de agua continentales son esenciales para el buen funcionamiento de nuestros ecosistemas, sin embargo, éstos se han visto afectados por estresores como el cambio climático y la eutrofización. El aumento de temperatura y nutrientes disueltos en el agua generan cambios significativos en la abundancia y composición del fitoplancton, principalmente en las poblaciones de cianobacterias las cuáles prosperan en estas condiciones generando así florecimientos algales (FA) que pueden ser nocivos para la salud del ecosistema. El estudio y monitoreo de estas poblaciones del fitoplancton es esencial para generar estrategias de control de la calidad del agua. Tradicionalmente el monitoreo de concentración de clorofila *in-situ* es utilizado para estudiar las poblaciones de algas. Sin embargo, este método presenta algunas limitaciones como sus altos costos y la inaccesibilidad a regiones de interés en los cuerpos de agua que genera una baja representatividad de las muestras. En este sentido las técnicas de teledetección, en conjunto con el desarrollo de nuevos sensores y nuevas metodologías para el procesamiento de datos, ofrecen una alternativa relativamente más accesible. Además, presenta alta periodicidad temporal y precisión para medir y monitorear indicadores de calidad del agua como lo es la concentración de clorofila. El área de estudio de este trabajo es lago-cráter de Santa María del Oro (SAMAO) en Nayarit, México. SAMAO presenta florecimientos algales de manera cíclica anual, el florecimiento y posterior decaimiento de las poblaciones de algas crea cambios de color en el agua, generalmente en la primera mitad del año. El objetivo de este trabajo es validar una estructura metodológica para el mapeo continuo de concentración de clorofila en SAMAO, para esto se evaluaron algoritmos de aprendizaje de máquina para estimación de clorofila empleando datos del sensor OLCI, y algoritmos de clasificación supervisada empleando una serie de datos del sensor MODIS. El capítulo 1 se presenta la introducción y estado del arte del desarrollo de algoritmos para estimar concentración de clorofila y clasificación. En el capítulo 2, se generó una base de datos de concentración de clorofila *in-situ* y reflectividades de OLCI, la cual se empleó para llevar a cabo la evaluación de diferentes algoritmos de estimación empleando diferentes estrategias en el tratamiento de la base de datos para mejorar la estimación. Todos los algoritmos de aprendizaje de máquina mejoraron la precisión de estimación en relación con el algoritmo OC_4 en la fase de entrenamiento, cumpliendo así con la hipótesis planteada. Los métodos *Kernel* entregaron las mayores mejorías con un 91.4% más respecto al OC_4. En el capítulo 3, se evaluaron diferentes algoritmos de clasificación supervisada para identificar los cambios ópticos en el lago usando datos usando datos de los productos MOD09GA/MYD09GA en el período de enero 2003 a diciembre 2020. El mejor algoritmo clasificador fue *Random Forest* con una precisión de 87.1%. Al aplicarlo a la serie de datos completa se encontró que mayo, abril y marzo son los meses con mayor presencia de cambios de color en el lago relacionados a FA, también que la mayor incidencia de florecimientos se da en la región noreste del lago y las mayores cantidades de eventos ocurrieron en los años 2011, 2008 y 2012 respectivamente. En el análisis de anomalías se encontró que el comportamiento interanual de los florecimientos puede atribuirse al fenómeno El Niño Oscilación del Sur con una correlación de -0.37.

Abstract

Inland water bodies are essential for the proper functioning of our ecosystems, however, they have been affected by stressors such as climate change and eutrophication. The increase in temperature and nutrients dissolved in the water generate significant changes in the abundance and composition of the phytoplankton, mainly in the populations of cyanobacteria which thrive in these conditions, thus generating algal blooms (AB) that can be harmful to the health of the ecosystem. The study and monitoring of these populations is essential to generate water quality control strategies. Traditionally, *in-situ* chlorophyll concentration monitoring is used to study these algae populations. However, it has some limitations such as its high costs and the inaccessibility to regions of interest in water bodies that generates a lack of representativity of the samples. In this sense, remote sensing techniques, together with the development of new sensors and new methodologies for data processing, offer a relatively cheaper alternative, with high temporal frequency and precision to measure and monitor water quality indicators such as the chlorophyll concentration. The study area in this work is the crater lake of Santa María del Oro (SAMAO) in Nayarit, Mexico. SAMAO presents algae blooms in an annual cyclical way, the blooming and subsequent decay of the algae populations creates color changes in the water, usually in the first half of the year. The objective of this work is to validate a methodological structure for the continuous mapping of chlorophyll concentration in SAMAO, for this, machine learning algorithms for chlorophyll estimation were evaluated using data from the OLCI sensor, and water color classification algorithms using a MODIS sensor data set. In chapter 1, an introduction and state-of-the-art development of algorithms for chlorophyll a estimation and classification is presented. In chapter 2, a database of in-situ chlorophyll concentration and OLCI reflectivities was generated, which was used to carry out the evaluation of five estimation algorithms using different strategies in the treatment of the database for improve estimation. All the machine learning algorithms improved the estimation accuracy in relation to the OC4 algorithm in the training phase, thus fulfilling the proposed hypothesis. Kernel methods delivered the greatest improvements with 91.4% more than OC4. In chapter 3, different supervised classification algorithms were evaluated to identify the changes in the red and near-infrared regions in the lake using data from the MOD09GA/MYD09GA products in the period from January 2003 to December 2020. The best classifying algorithm was Random Forest with an accuracy of 87.1%, when applied to the complete data series it was found that May, April and March are the months with the greatest presence of color changes in the lake related to AF, also that the The highest incidence of blooms occurs in the northeast region of the lake and the highest number of events occurred in the years 2011, 2008 and 2012 respectively. In the analysis of anomalies, it was found that the interannual behavior of the blooms can be attributed to the El Niño Southern Oscillation phenomenon with a correlation of -0.37.

Capítulo I

Introducción general

Las aguas continentales incluyen los lagos, ríos, arroyos, canales y presas (Fisheries, 2011), representan aproximadamente el 0.01% del volumen total del agua en la Tierra (Stiassny *et al.,* 1996) y cubren solo el 0.8% de la superficie terrestre (Gleick, 1996). A pesar de constituir una pequeña fracción del agua total del planeta, los cuerpos de agua continentales (CAC) tienen funciones muy importantes para el ambiente ya que proveen de hábitat a aproximadamente un tercio de las especies de vertebrados (Dudgeon *et al.,* 2006) y albergan un ~40% de la diversidad de peces del planeta, además de anfibios, reptiles y mamíferos (Lundberg *et al.,*2000). Las aguas continentales incluyen los lagos, ríos, arroyos, canales y presas (Fisheries, 2011), representan aproximadamente el 0.01% del volumen total del agua en la Tierra (Stiassny *et al.,* 1996) y cubren solo el 0.8% de la superficie terrestre (Gleick, 1996). A pesar de constituir una pequeña fracción del agua total del planeta, los cuerpos de agua continentales (CAC) tienen funciones muy importantes para el ambiente ya que proveen de hábitat a aproximadamente un tercio de las especies de vertebrados (Dudgeon *et al.,* 2006) y albergan un ~40% de la diversidad de peces del planeta, además de anfibios, reptiles y mamíferos (Lundberg *et al.,* 2000).

Los CAC son también un componente esencial del ciclo del carbono (Tranvik *et al.,* 2009), del nitrógeno (Moss, 2012) y en general de nutrientes (Carpenter *et al.,* 2011). Afectan el clima regional a través del intercambio de calor y agua con la atmósfera (Krinner, 2003). Además, son aprovechados por los humanos de múltiples maneras, por ejemplo, en la

producción de energía, pesca, actividades recreativas, como un medio para transportarse y más importante, para consumo humano directo y para la irrigación de cultivos (Lynch *et al.,* 2016; Dörnhöfer & Oppelt, 2016; Stendera *et al.,* 2012). A pesar de lo anterior, estos son los ecosistemas con la mayor tasa de peligro de extinción en el mundo (Dudgeon *et al.,*2006). El declive de la biodiversidad es más alto en los CAC es más alta que en cualquier ecosistema terrestre (Sala *et al.,*2000) debido a que factores como el cambio climático y la eutrofización amenazan sus funciones ecológicas, principalmente estresores (Glibert, 2020; Dörnhöfer & Oppelt, 2016).

El proceso de eutrofización en cuerpos de agua continentales se entendía como un fenómeno natural inducido por procesos autóctonos (Hynes, 1969). Sin embargo, la eutrofización cultural, inducida por actividades antropogénicas, es una condición que domina en las aguas superficiales de todos los continentes (Harper *et al.,*1992) y, al igual que el cambio climático, se ha convertido en un problema ambiental global que se pronostica se intensificará en las siguientes décadas debido principalmente al crecimiento demográfico, el aumento de temperatura en el planeta y al cambio de uso de suelos (Glibert, 2020; O'neil *et al.,* 2012; Carpenter, 2005). La eutrofización de los cuerpos de agua implica el enriquecimiento de nutrientes (principalmente nitrógeno y fósforo), seguido a menudo de cambios significativos en la abundancia y composición del fitoplancton, generalmente en las poblaciones de cianobacterias (Steinberg & Hartmann, 1988), las cuales no solo presentan diversas estrategias fisiológicas que les brindan ventajas competitivas sobre otras especies de algas, sino que también tienen la característica de producir metabolitos secundarios biológicamente activos que se acumulan en el citoplasma denominados cianotoxinas, (Uriza *et al.,*2017; Paerl & Millie,

1996) esto las hace ser las principales responsables de eventos de intoxicación en aguas dulces (Pizzolon, 1996).

A los crecimientos poblacionales de cianobacterias se les denomina florecimientos algales nocivos (FAN's). Tras el crecimiento exponencial de la población, la muerte de estos organismos puede llevar al agotamiento del oxígeno disuelto en el agua, lo que a su vez puede provocar problemas secundarios como mortalidad de peces y liberación de sustancias tóxicas y fosfatos que previamente se unieron a sedimentos oxidados. Los fosfatos liberados de los sedimentos aceleran la eutrofización, cerrando así un ciclo de retroalimentación positiva (Jamie Bartram 2015) que degrada cada vez más la calidad del agua. Se han detectado y documentado FAN's desde 1980, sin embargo, se conocen informes de casos de intoxicación de hace más de mil años (Bartram *et al.,*1999). Actualmente, es inequívocamente reconocido que la expansión global de FAN's continúa con cada vez mayor abundancia, frecuencia y extensión geográfica, con nuevas especies reportándose en nuevas áreas (Glibert, 2020). En 1998, la Organización Mundial de la Salud (OMS) estableció como valor provisional de referencia 1 mg/L de microcistinas como nivel máximo aceptable en aguas de abastecimiento público para el consumo oral diario (Uriza *et al.,* 2017). Países como Brasil (Azevedo *et al.,* 2002), Canadá (Health Canada, 2003) y Australia (NHMRC, 2006) incluyeron a las cianotoxinas en su legislación respecto a calidad de agua. A pesar de su impacto, son pocos los países que han puesto atención al tema. Es apremiante implementar políticas ambientales y en materia de salud pública que atiendan esta problemática; para ello es necesario estudiar la presencia de los crecimientos algales, conocer las concentraciones, su temporalidad e identificar los efectos negativos para que puedan planearse acciones de control y manejo de cuencas.

Los FAN's se monitorean usando medidas de la biomasa junto con la examinación de especies presentes. Las algas unicelulares que conforman el fitoplancton, así como las cianobacterias, contienen cloroplastos que absorben y usan la luz que atraviesa el agua para fijar carbono en forma de carbohidratos (Suthers *et al.,* 2019). Entre los pigmentos presentes en los cloroplastos, la clorofila-a (*Chl-a*) es la más común en todo el fitoplancton (Han & Jordan, 2005). Ésta es una molécula ópticamente activa en cuerpos de agua, actúa como índice del estado de eutrofización, por lo que es una medida ampliamente utilizada para determinar la biomasa de algas (Zurawell, 2015). El tradicional monitoreo *in-situ* de la calidad de agua es crucial para cualquier esfuerzo de producir información en apoyo a la conservación del agua y en la toma de decisiones (Duan *et al.,*2007), sin embargo, se vuelve complejo implementar un monitoreo basado en medidas *in-situ* que permita observar de forma continua los cambios de la calidad del agua debido a los altos costes, la disponibilidad de tiempo, la inaccesibilidad a las áreas de interés del lago o el tamaño y número de cuerpos de agua que se desea estudiar; en consecuencia, éstos muestreos se concentran sólo en áreas desarrolladas o fácilmente accesibles, lo que provoca una baja representatividad debido a la irregularidad espacial y la toma no aleatoria de las muestras (Masocha *et al.,* 2018). En este aspecto, las técnicas de *sensores remotos* (SR) han sido utilizadas para superar las limitaciones de las mediciones tradicionales (Torbick *et al.,* 2008).

Los SR ofrecen un método relativamente más barato, con alta periodicidad temporal y elevada precisión para medir y monitorear indicadores de calidad del agua (Guan *et al.,* 2011; Torbick *et al.,* 2013; Dube *et al.,* 2015; Masocha *et al.,* 2018). La estimación remota de los constituyentes del agua se basa en la relación entre la energía incidente y las propiedades ópticas inherentes del cuerpo de agua (Gitelson *et al.,* 2009); esto es posible con proporciones adecuadas

de bandas espectrales y sus combinaciones, este método relaciona las mediciones de reflectancia con las concentraciones de *Chl-a* a partir de modelos de reflectividad y ecuaciones empíricas (Duan *et al.,* 2010). Los datos satelitales son medidas del flujo radiante (radiancia e irradiancia) que llega y sale de la Tierra, la cual varía según la latitud, temporada del año y hora del día (Brezonik *et al.*, 2005). De forma que los algoritmos son desarrollados para cuerpos de agua y temporadas particulares, por lo tanto, no pueden ser usados en todas partes ni en todas las temporadas del año (Kutser *et al.,*2001). De modo que, esfuerzos significativos se han empleado en el desarrollo de novedosos algoritmos que estimen con mayor precisión la concentración de clorofila en cuerpos de agua continentales relacionando las mediciones de reflectancia con las concentraciones de *Chl-a in-situ* (Singh *et al.,* 2014). Con el avance tecnológico y de capacidad computacional las técnicas de aprendizaje de máquina van adquiriendo mayor popularidad en el desarrollo de algoritmos de estimación de parámetros biofísicos (Cao *et al.,* 2020; Blix *et al.,* 2019; Keller *et al.,* 2018; Verrelst *et al.,* 2012), pues en general son capaces de manejar problemas complejos sin ningún conocimiento de dominio (o sea, sin necesidad de un modelo físico) cuando se tienen suficientes datos de entrada.

Marco teórico

Bio-óptica

La bio-óptica es la rama de la ciencia que estudia la relación entre la luz y los componentes del agua. El espectro electromagnético (EM) hace referencia a diferentes tipos de radiación, la cual es energía que viaja en forma de ondas (Figura 1). Cada onda electromagnética tiene una longitud particular, donde la radiación con longitudes de onda larga contiene menos energía que aquella con longitudes de onda corta (Lipson, 2010).

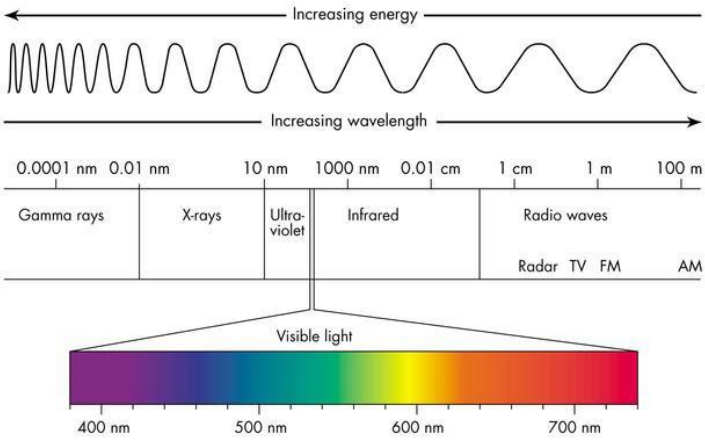


Figura 1. Esquema general del espectro electromagnético (O'Connor, 2013).

El intervalo del espectro electromagnético entre los 400 a 700 nm constituye la luz fotosintéticamente activa (PAR, acrónimo del inglés Photosynthetically Available Radiation), dichas longitudes de onda son utilizadas por el fitoplancton que absorbe fotones para la fotosíntesis (Matthews, 2017. Este intervalo también es la única parte del espectro electromagnético que puede ser detectada por el ojo humano. Los fotones de la luz solar actúan como ondas, sin embargo, también pueden actuar como partículas (OpenStax College, 2013). La bio-óptica nos ayuda a entender los factores que contribuyen a regular la transferencia de la luz dentro del agua, estos factores están relacionados a procesos biogeoquímicos, y en especial con la productividad primaria fitoplanctónica y el ciclo global del carbono (Ramus 1995). El agua natural es una mezcla de materia disuelta y partículas en suspensión, estos solutos y partículas son ópticamente significativos y muy variables en tipo y concentración. Debido a esto, las propiedades ópticas del agua muestran grandes variaciones temporales y espaciales y rara vez se asemejan a las del agua pura. Las propiedades ópticas del agua se dividen en dos clases: inherentes y aparentes (Mobley 1994).

Calidad del agua y clorofila

Las propiedades de calidad del agua como, el total de sólidos suspendidos, porcentaje de saturación de oxígeno disuelto, y la turbidez, se utilizan como indicadores primarios para evaluar la viabilidad ambiental de las aguas lacustres y costeras, lo cual permite a las instituciones ambientales, como la Comisión Nacional del Agua, orientar la gestión de recursos y las decisiones de seguridad pública. La calidad del agua puede ser medida por una sola variable o por la combinación de más de cien. Para la mayoría de los propósitos, puede ser descrita por un poco menos de 20 características físicas, químicas y biológicas. Las variables elegidas en el programa de monitoreo dependerán de los objetivos y los usos existentes y previstos. Hay tres medios principales que se pueden utilizar para el monitoreo de cuerpos acuáticos: agua, partículas y organismos vivos (Bartram & Ballance, 1996). Las partículas suspendidas, por ejemplo, materia suspendida total, la profundidad de Secchi y las concentraciones de nutrientes se pueden determinar mediante análisis físicos y químicos; mientras que para los organismos vivos se pueden utilizar de diferentes aproximaciones. La medición biológica más común de las muestras de agua superficiales es la determinación de la *Chl-a*, cuya concentración es utilizada como un indicador de la biomasa de fitoplancton. Las mediciones de *Chl-a* son posiblemente el descriptor ambiental más completo pues son útiles para evaluar la eutrofización en lagos, embalses y grandes ríos.

Hay cinco tipos de clorofilas caracterizadas estructuralmente: *a, b, c, d y f* (Airs *et al.,*2014). Éstas exhiben diferentes máximos de absorción debido a modificaciones relativamente menores de su estructura química (Chen *et al.,*2012). En las plantas, incluidas las algas, la *Chl-a* y *Chl-b* son los principales pigmentos fotosintéticos.re La *Chl-a* constituye de

manera aproximada el 75% de toda la clorofila de las algas verdes (Reol, 2003). Este pigmento absorbe radiación sobre todo en la parte del rojo (650-700 nm), violeta (400 nm) y azul (450-490 nm) y refleja en el verde (490-530 nm) (Reol, 2003) del espectro electromagnético. A la forma en que un material refleja, emite o absorbe energía se le conoce como firma espectral. De modo que la firma espectral permite identificar y discriminar diferentes objetos a partir de la señal registrada por un sensor en las diferentes regiones del EM.

La *Chl-a* es el principal pigmento fotosintético en el fitoplancton, el cuál forma la base de la red alimentaria en cuerpos de agua y oceánica, y es responsables de aproximadamente la mitad de la fotosíntesis global (Behrenfeld & Falkowski, 1997). Además de los métodos químicos, la *Chl-a* puede ser estimada con datos de *reflectividad medida con sensores remotos* (Rrs) usando diferentes algoritmos para obtener mejores estimaciones de *Chl-a* a través de un rango amplio de tipos de agua.

Clasificación de tipos de agua

Por muchas décadas se ha empleado la clasificación de aguas en Caso I y Caso II, las primeras hacen referencia a aguas oligotróficas dónde las propiedades ópticas dependen principalmente del fitoplancton, por ejemplo, las aguas oceánicas. Las aguas continentales y costeras se incluyen en las aguas Caso II, éstas muestran una gran variabilidad en sus propiedades ópticas pues tienen el efecto continental del arrastre de sedimentos, nutrientes y materia orgánica y la resuspensión de los sedimentos del fondo. Sin embargo, el icónico sistema de clasificación caso 1/caso 2 que ha predominado durante las últimas décadas no es en realidad un sistema de clasificación objetivo (Moore *et al.,*2014), especialmente para el desarrollo de algoritmos de estimación de parámetros biofísicos por medio de datos satelitales, donde se

requiere un enfoque más específico de clasificación de aguas. Es ya sabido que los algoritmos que emplean bandas del verde/azul para estimar *Chl-a* funcionan adecuadamente con un margen de error aceptable en aguas oceánicas caracterizadas por propiedades ópticas simples. Sin embargo, los retos siguen en las aguas continentales y costeras donde las propiedades ópticas son complejas y variables debido a las señales ópticas mezcladas del material suspendido en el agua (Cui *et al.,*2020).

Muchos algoritmos con bandas del rojo e infrarrojo se han desarrollado y han demostrado generar estimaciones precisas de *Chl-a* en aguas continentales y costeras, sin embargo, la amplia gama de posibles combinaciones y la composición del material suspendido que se encuentran dentro y entre los sistemas acuáticos Caso II desafía la aplicabilidad de las técnicas de Observación de la Tierra. Podemos decir que, las clases ópticas, y por lo tanto la precisión de los algoritmos, varían dentro de los entornos costeros y lacustres para aguas que podrían denominarse colectivamente como Caso II (Moore *et al.,*2014). Es ampliamente aceptado que no es factible aplicar un algoritmo bio-óptico universal a diferentes cuerpos de agua (Moore, Campbell, and Feng 2001) o incluso al mismo cuerpo de agua en diferentes tiempos. Los algoritmos bio-ópticos se desempeñan mejor en ciertas condiciones y peor en condiciones diferentes, por lo que se hace necesario un esquema de clasificación que pueda diferenciar el entorno y elegir el algoritmo más apropiado para las condiciones ambientales dadas (Moore *et al.,*2014).

La pre-clasificación de la reflectancia satelital en varios tipos de agua tiene como fundamento las similitudes y diferencias en las características de la forma de su espectro, por lo que diferentes coeficientes de regresión y una óptima combinación de bandas podría usarse para

tipos de agua que comparten propiedades o atributos en común (Shi *et al.,*2013). Este concepto de clasificar primero los tipos ópticos de agua en función de las características de reflectancia y posteriormente desarrollar algoritmos específicos para cada tipo de agua es un esquema válido para reducir errores en la estimación de clorofila en cuerpos de agua continentales ópticamente complejos (Uudeberg *et al.,*2019). Esta metodología es una solución que puede resultar útil para comprender las variaciones temporales y espaciales de tipos de agua y para planificar programas de seguimiento de cuerpos de agua lacustre.

Corrección atmosférica en cuerpos de agua continentales

La señal recibida por un sensor pasivo viaja a través de la atmósfera dos veces, del Sol a la superficie terrestre, y de ahí de vuelta hacia el sensor. En este proceso la luz recibida por el sensor se ve afectada por absorción y dispersión provocada por las moléculas gaseosas y partículas de materia de la atmósfera. La corrección atmosférica es el proceso de corregir estos efectos atmosféricos para obtener la reflectancia de un objetivo de la superficie terrestre a partir de la radiancia medida por el sensor. El efecto de la atmósfera en la radiancia recibida por SR es significativamente mayor cuando se trata de cuerpos de agua, ya que el agua absorbe mucho y solo contribuye con el 20% o menos del total de radiancia que llega al sensor (Moses *et al.,*2017). Corregir estos efectos es un prerrequisito esencial para obtener estimaciones acertadas de radiancia, la cual es base para derivar estimaciones cuantitativas de parámetros biofísicos mediante SR. Desde el lanzamiento del primer sensor para color del océano, el Coastal Zone Color Scanner en 1978, se han desarrollado numerosos algoritmos para corregir los efectos de la atmósfera en datos espectrales medidos en aguas oceánicas. Sin embargo, aplicar estos algoritmos a cuerpos de agua continentales no es tan sencillo debido a: 1) la proximidad a

fuentes terrestres de contaminación atmosférica, lo que da como resultado una atmósfera ópticamente heterogénea que es difícil de modelar, 2) efectos de adyacencia de píxeles terrestres vecinos, efecto especialmente significativo en los casos de una topografía elevada y ondulada alrededor del cuerpo de agua, 3) reflectancia no despreciable del agua en la región del infrarrojo cercano debido a las altas concentraciones de sedimentos en las aguas continentales causada por descargas agrícolas e industriales de fuentes terrestres, escorrentías superficiales y subterráneas, resuspensión de sedimentos impulsada por el viento, y afluencia de sedimentos por deslizamientos de tierra y erosión de la costa. Todo esto dificulta estimar con precisión y eliminar el efecto de la contribución de aerosoles atmosféricos en la señal recibida. 4) Variaciones en la altitud de la superficie de las aguas continentales desde el nivel medio del mar, lo que introduce incertidumbres en las estimaciones del contenido de aerosoles en la columna atmosférica sobre el agua.

Modelos para estimar parámetros biofísicos

Los satélites ópticos de observación terrestre permiten obtener y monitorear variables bio-geofísicas como la clorofila del fitoplancton. La cuantificación de parámetros biofísicos en la superficie terrestre con sensores remotos siempre recae en un modelo numérico que permita la interpretación de las observaciones espectrales y su traducción a variables bio-geofísicas de la superficie terrestre (Dekker *et al.,*1996). Las determinaciones de variables bio-geofísicas están agrupados típicamente en dos categorías: a) la categoría estadística o dirigidos por la variable, b) La categoría física o dirigidos por los datos radiométricos. Ambas categorías del método se dividen a su vez en subcategorías y combinaciones de las mismas. Por ello, es necesario realizar una categorización sistemática de estos métodos ya que está aumentando el

número de elementos de ambas categorías (Verrelst *et al.,*2015). Así, los métodos de determinación de parámetros biofísicos mediante SR se pueden categorizar en: 1) modelos de regresiones paramétricas 2) modelos de regresiones no paramétricas 3) modelos basados en la física 4) métodos híbridos (Baret & Buis, 2008). A continuación, se describen brevemente los métodos utilizados para la determinación de las características bio-geofísicas.

Métodos de regresión paramétrica: asumen una relación explícita entre las observaciones espectrales y la variable bio-geofísica específica (figura 2). Así las expresiones paramétricas explícitas son construidas. Típicamente confiando en el conocimiento físico o estadístico de la variable y su respuesta espectral. Generalmente la formulación aritmética de bandas se define y después se relaciona con la variable de interés basada en una función de ajuste.

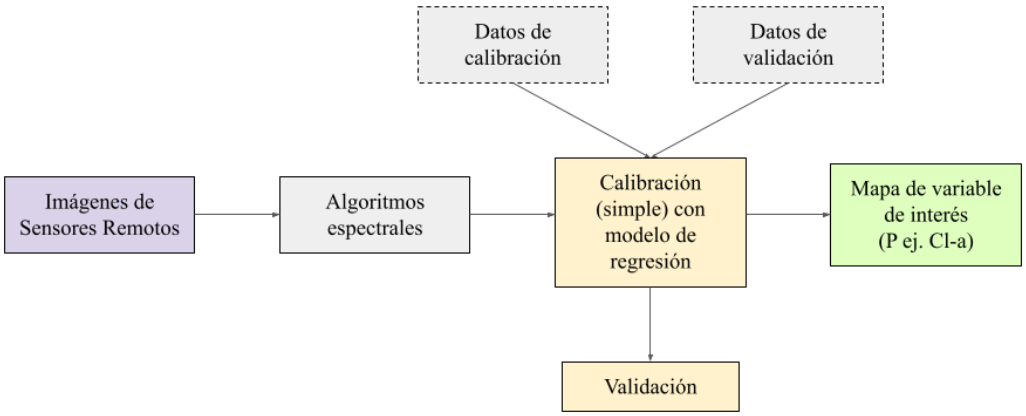


Figura 2. Diagrama de flujo del procedimiento general de los modelos de regresión paramétricos.

Métodos de regresión no paramétrica: definen directamente la función de regresión de acuerdo a información espectral por sensores remotos (figura 3). Por lo tanto, en contraste con los métodos de regresión paramétrica, no se debe hacer una elección explícita en las relaciones

de bandas espectrales, sus transformaciones o funciones de ajuste. Los métodos de regresión no paramétrica se pueden dividir en lineales o no lineales. Los algoritmos de regresión no-paramétricos no-lineares se conocen como algoritmos de regresión de aprendizaje de máquina.

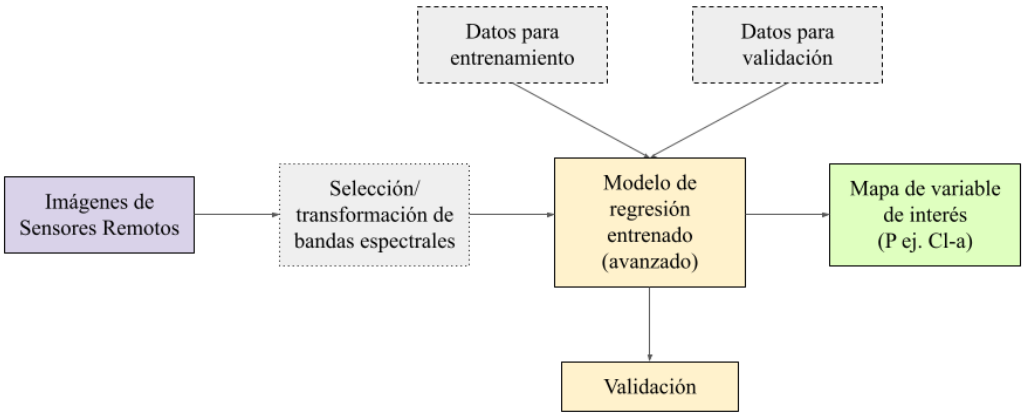


Figura 3. Diagrama de flujo del procedimiento general de los modelos de regresión no paramétricos.

Métodos basados en física: son aplicaciones de las leyes físicas estableciendo relaciones causa-efecto. Las variables del modelo son inferidas basadas en conocimiento específico, típicamente obtenidos con funciones de transferencia radiativa.

Métodos híbridos: combina elementos de los métodos estadísticas no paramétricas y de los basados en física. Hacen uso de propiedades genéricas de métodos basados en física combinados con la flexibilidad y eficiencia computacional de los métodos de regresión no-paramétricos no-lineares.

Aprendizaje de máquina

El aprendizaje de máquina es el campo de estudio que da a las computadoras la habilidad de aprender sin ser programadas explícitamente (Samuel 1959). Para resolver problemas con una computadora, necesitamos un algoritmo. Un algoritmo es una secuencia de instrucciones que deberían llevarse a cabo para transformar las 'entradas' (p ej. datos espectrales) en 'salidas' (p ej. concentración de *Chl-a*). Puede haber diferentes algoritmos que resuelvan el mismo problema, pero lo que nos interesa es encontrar el más eficiente, que requiera una menor cantidad de instrucciones o memoria, o ambos. Sin embargo, para algunos problemas no tenemos algoritmos, sabemos cuáles son las entradas y cuáles esperamos que sean las salidas, pero no sabemos cómo transformar las entradas en salidas. Para estos casos buscamos que la computadora (máquina) extraiga automáticamente el algoritmo para esta tarea reconociendo patrones en los datos de entrada que se emplean para su entrenamiento y "aprendiendo" de éstos (Alpaydin 2020). Con el aprendizaje de máquina compensamos con datos lo que carecemos en conocimiento. El aprendizaje es un dominio muy amplio, en consecuencia, el campo del aprendizaje de máquina se ha ramificado en diferentes subcampos que hacen frente a diferentes tipos de tareas de aprendizaje. Una de las principales divisiones del aprendizaje de máquina es la de **aprendizaje supervisado y el no supervisado**. En el supervisado el aprendiz (la máquina) recibe un conjunto de datos con etiqueta, y en base a esto debe encontrar una regla para etiquetar los siguientes datos de entrada nuevos que se le den. Por otro lado, en el no supervisado, todo lo que recibe la máquina es un conjunto amplio de datos sin etiquetar y ésta por sí misma debe procesar los datos de entrada con el objetivo de generar un resumen o una versión comprimida de ellos de acuerdo a sus características (Shalev-Shwartz & Ben-David, 2014). El 'clustering' de un conjunto de datos en subconjuntos de objetos similares es un ejemplo típico.

Verrelst *et al.,* (2015) recopilaron algunos ejemplos de algoritmos de aprendizaje de máquina, los cuales se muestran a continuación:

- Árboles de decisión: Un árbol de decisión se basa en un conjunto de nodos conectados jerárquicamente. Cada nodo representa una decisión lineal basada en una característica de entrada específica.
    a. Bagging decision trees (Breiman 1996)
    b. Random Forests (RF) (Breiman 2001)
- Redes neuronales artificiales: son básicamente una función no lineal puntual (por ejemplo, una función sigmoidea o gaussiana) aplicada a la salida de una regresión lineal. Se llaman de esta forma ya que en esencia son estructuras de neuronas artificiales en capas completamente conectadas.
- Métodos Kernel: deben su nombre al uso de funciones de núcleo. Los núcleos cuantifican las similitudes entre las muestras de entrada de un conjunto de datos (Shawe-Taylor, Cristianini, *et al.,*2004).
    a. Support vector machines (SVM's)
    b. Kernel ridge regression (KRR)
    c. Relevance vector machines (RVM)
    d. Gaussian processes regression (GPR)
- Redes Bayesianas: son una clase de modelos probabilísticos, se caracterizan por estructuras gráficas que representan información sobre dominios de incertidumbre

(Cooper&Herskovits, 1992). Las Redes Bayesianas se estructuran utilizando gráficos acíclicos dirigidos. Cada nodo en el gráfico representa una variable aleatoria, mientras que los bordes de los nodos conectan las dependencias probabilísticas entre variables.

Área de estudio

El lago de Santa María del Oro (SAMAO) se encuentra a 730 m.s.n.m. dentro de una estructura volcánica, en el lado noroeste del Cinturón Volcánico Mexicano (Armienta *et al.,*2008) (figura *4*). Es un lago de edad pleistocénica con carácter endorreico (Sosa-Nájera *et al.,*2010), diámetro aproximado de 2 kilómetros, una superficie de 3.7 km² y una profundidad máxima de 58 m. Localizando su coordenada central en 21°22’58” N, 104°34’48” W. Durante el período de 1981 al 2010 SAMAO registró en la estación meteorológica de Cerro Blanco, número 18005 del Servicio Meteorológico Nacional, una temperatura media anual de 20.9 °C, siendo mayo el mes más cálido con una temperatura máxima mensual registrada de 35.6 ° C, mientras que en enero se presenta la mínima de 3.5 °C. La temporada de lluvias va de junio a octubre y su precipitación media anual va de 1214 a 1600 mm (Sosa-Nájera *et al.,*2010, Serrano *et al.,*2002). SAMAO pertenece a la región hidrográfica Lerma-Santiago, en la cuenca Santiago-Aguamilpa y dentro de la subcuenca del río Mojarras. Esta subcuenca se caracteriza por la presencia de los arroyos Zapotanito, La Cofradía y El Buruato, los cuales drenan sus aguas en el río Grande Santiago. Tales afluentes tienen importantes aportes de agua a los mantos acuíferos y cuerpos de agua superficiales de la zona, como lo es la Laguna Santa María del Oro (Arreola&Morales, n.d.).

Los principales cultivos de la zona son el maíz y el frijol, aunque existen pequeñas zonas de plantación de mango y plátano en huertas, por lo que hay cultivos anuales, cultivos perennes

y cultivos semiperennes (Arreola & Morales, n.d.). SAMAO es una zona importante para el abastecimiento de agua, acuicultura, agricultura y turismo (Serrano *et al.,* 2002). A pesar de la falta de conocimientos sobre la flora y fauna acuática, se estima que la región de los lagos volcánicos, (incluyendo a Tepetiltic y San Pedro) es una zona de endemismo concentrado y de elevada biodiversidad (CONABIO, 2018).

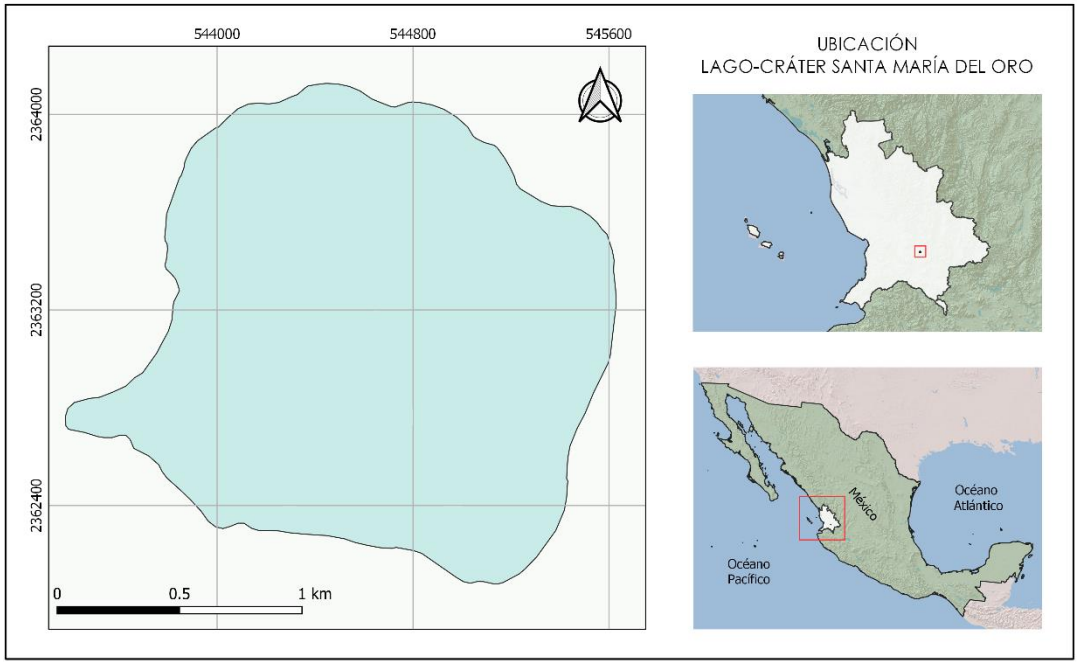


Figura 4. Mapa de ubicación del lago-cráter de Santa María del Oro.

Un rasgo sobresaliente de la ecología de SAMAO es que presenta Florecimientos Algales Nocivos (FAN) de Cianobacterias (Salazar-Alcaraz 2018) en forma cíclica en escala anuales, y se presentan generalmente entre los meses de febrero-marzo. Se han identificado tres especies de cianobacterias formadoras de estos florecimientos: *Limnoraphis robusta, Microcystis aeruginosa, y Microcystis smithii (Ochoa-Zamora 2018).*

Justificación

Los recursos de los cuerpos de agua continentales son de suma importancia para la salud de las comunidades. Es vital estudiar estas aguas de forma sistemática y con una perspectiva a largo plazo para caracterizar la variabilidad de sus propiedades bio-ópticas y biogeoquímicas, y de esta forma comprender sus impactos en la calidad del agua.

Hipótesis

Dado que Sentinel-3 fue diseñado para el monitoreo global de la calidad del agua, se espera que emplear sus datos para el desarrollo de un algoritmo de estimación de clorofila aplicando métodos de aprendizaje de máquina mejorará la precisión de la estimación en relación con los algoritmos operativos actualmente.

Objetivo

Validar una estructura metodológica para el mapeo continuo de concentración de clorofila en cuerpos de agua continentales (caso de estudio SAMAO), empleando algoritmos de aprendizaje de máquina y datos de sensores remotos.

Objetivos específicos

- Generar una base de datos de medidas de reflectividad del sensor OLCI (Sentinel-3) y concentración de clorofila superficial *in-situ* por técnicas de fluorometría durante un año.
- Evaluar al menos tres algoritmos de estimación de concentración de *Chl-a* usando técnicas de aprendizaje de máquina implementadas en el software ARTMO.
- Caracterizar la dinámica espacio-temporal de los florecimientos algales del lago-cráter de Santa María del Oro a partir de datos de reflectividad del sensor MODIS (MOD09GA, MYD09GA) durante los años 2003-2020.

Capítulo II

Evaluación de algoritmos de regresión para *Chl-a*

Resumen del capítulo

Se generó una base de datos (BD) para vincular mediciones de la concentración de *Chl-a in-situ* con estimaciones de reflectividad del sensor OLCI. Se realizaron 7 muestreos en SAMAO durante el año 2020 para estimar *Chl-a,* las mayores concentraciones de *Chl-a* se obtuvieron en los muestreos de marzo y mayo con valores por encima de 41.7µg/L, la menor concentración se obtuvo en febrero con valores de 0.9 µg/L. Se realizó un análisis de correlación entre las bandas de OLCI para determinar qué bandas aportan mayor información espectral adecuada para el entrenamiento del algoritmo de estimación. Mediante K-means se clasificaron los perfiles espectrales de SAMAO en tipos agua, dónde la mayor separabilidad se obtuvo con 2 clases de acuerdo al coeficiente Silhouette, sin embargo, se encontró que no existe un patrón claro que permita asociar rangos específicos de *Chl-a* a alguna de las clases. A continuación, se empleó la BD obtenida para entrenar y evaluar 5 algoritmos de aprendizaje de máquina y compararlos con el algoritmo operativo OC_4. Todos los algoritmos de aprendizaje mejoraron la estimación de concentración de *Chl-a* usando la validación Leave-One-Out: Mínimos cuadrados parciales mejoró la estimación con un RMSE de 10.05 en comparación con OC_4, que obtuvo un RMSE de 10.8. Mientras que en la fase de entrenamiento, los métodos Kernel rigde regresion y Gaussian Process obtuvieron un RMSE de 0.9 y 1.1 respectivamente. Se evaluaron tres estrategias para mejorar la estimación y la validación de los modelos, donde los mejores resultados se obtuvieron utilizando la base de datos completa sin hacer una clasificación óptica de la respuesta espectral de SAMAO, a diferencia de la fase de entrenamiento de los modelos donde hubo ventajas empleando la clasificación óptica. Nuestros resultados indican que es necesario mejorar el número de muestras utilizadas en la base de datos para optimizar la representabilidad espacial de las muestras de concentración de *Chl-a* colectadas *in-situ*.

Introducción

La eutrofización de los cuerpos de agua implica el enriquecimiento de nutrientes seguido a menudo de cambios significativos en la abundancia del fitoplancton, generalmente en las poblaciones de cianobacterias (Tomaselli *et al.,* 2004). Si bien, la eutrofización de los cuerpos de agua continentales se entendía como un fenómeno natural inducido por procesos autóctonos a una escala de miles de años (Larkin *et al.,*1969), la eutrofización cultural, inducida por actividades antrópicas, ha acelerado este proceso creando un problema ambiental global. Se pronóstica que este fenómeno se intensificará en las siguientes décadas debido principalmente al aumento demográfico y el cambio de uso de suelos (Carpenter, 2005).

Las algas son organismos unicelulares que conforman el fitoplancton, a través de los cloroplastos estas absorben y usan la luz que atraviesa el agua para fijar carbono en forma de carbohidratos (Suthers *et al.,* 2019). Entre los pigmentos presentes en los cloroplastos, *Chl-a* es el más común en todos los grupos de algas por lo que es uno de los principales indicadores de biomasa, productividad primaria y del estado de eutrofización en cuerpos de agua continentales (Han & Jordan, 2005).

El monitoreo de la concentración de *Chl-a* es clave para evaluar la calidad del agua en lagos y embalses (Duan *et al.,* 2007). Al realizar las campañas *in-situ* la muestra se concentra en áreas desarrolladas o fácilmente accesibles, lo que crea una irregularidad espacial y muestras no aleatorias (Masocha *et al.,* 2018). La *Chl-a* al ser una molécula ópticamente activa con características conocidas de dispersión y absorción de luz, puede ser detectada y monitoreada a través de sensores remotos, los cuales permiten un monitoreo temporal más frecuente y sinópticas en los CAC (Nas *et al. ,*2010; Torbick *et al.,* 2013; Dube *et al.,* 2015; Masocha *et al.,*

2018). Es por esto, que se realizan esfuerzos significativos en el desarrollo de nuevos algoritmos que estimen con mayor precisión la concentración de *Chl-a* en CAC (Singh *et al.,* 2014).

La determinación de *Chl-a* con sensores remotos se logra a partir de la aplicación de algoritmos que integran datos limnológicos colectados *in-situ* con mediciones de reflectividad obtenidas a través de sensores en satélites puestos en órbita (Schowengerdt, 2006). El monitoreo de *Chl-a* mediante sensores remotos satelitales tiene una historia de más de 30 años (Comiso *et al.,*1993). Los sensores de la primera y segunda generación, como el Coastal Zone Color Scanner (CZCS) (Hovis, 1980) y el Sea-viewing Wide Field-of-view Sensor (SeaWiFS) (Hooker, 1992), fueron desarrollados para monitorear el color del océano aplicando algoritmos que emplean relaciones empíricas con las bandas del azul y verde (O'Reilly *et al.,*1998). Desde entonces, se ha desarrollado una amplia gama de sensores ópticos con diferentes resoluciones espectrales, espaciales y temporales (Xing *et al.,*2007), y se planea poner en órbita más misiones que traerán consigo un flujo de datos sin precedentes. Por lo anterior es indispensable desarrollar técnicas de procesamiento eficientes que permitan la cuantificación espacial y temporal explícita de forma rápida y continua (Verrelst *et al.,* 2012) por lo que, se han estado explorando nuevos métodos de estimación para *Chl-a* que además permitan usar los datos de las misiones anteriores.

Los métodos de determinación de parámetros biofísicos mediante sensores remotos se pueden categorizar en: 1) modelos de regresiones paramétricas, 2) modelos de regresiones no paramétricas, 3) modelos basados en leyes físicas, 4) métodos híbridos (Baret & Buis, 2008). Los modelos de regresión no paramétricos no requieren una selección explícita de bandas espectrales o sus transformaciones. En contraste con los paramétricos estos modelos son

flexibles, capaces de combinar diferentes características de la estructura de datos de manera no lineal de acuerdo con lo que se desea obtener. Los métodos no paramétricos se han vuelto más frecuentes en la literatura reciente (Su *et al.,*2021; Jeong *et al.,*2022; *Lin et al.,* 2023), estos métodos se pueden subdividir en modelos no paramétricos lineares y no lineares (Verrelst *et al.,*2015). Los modelos no paramétricos no lineares, también llamados algoritmos de regresión de aprendizaje de maquina (MLRAs, por las siglas en inglés de Machine Learning Regression Algorithms) se han estado desarrollando desde hace varias décadas (DeFries & Chan, 2000), éstos asumen que las relaciones entre las características de los datos no son explícitas por lo que permiten modelar datos reales en forma "natural", permitiendo también la incorporación de diferentes tipos de datos en el análisis (Verrelst *et al.,* 2015). Entre los algoritmos no paramétricos no lineares encontramos:

- Random Forest (RF) (Breiman, 2001), en Shen *et al.,* (2022) se demuestra la capacidad de RF para obtener valores precisos de *Chl-a* en CAC eutrofizados empleando datos OLCI. Se evaluó su desempeño junto a los algoritmos *extreme gradient boosting* (XGBoost), *deep neural network* (DNN) y *support vector regression* (SVR), donde RF mostró tener menor sensibilidad a las incertidumbres de la corrección atmosférica, obteniendo los mejores resultados de la evaluación.

- Redes neuronales (NN) (Jain *et al.,* 1996), Vilas *et al.,* (2011) empleó datos MERIS, precursor de OLCI, para obtener mediciones de *Chl-a* en cuatro embalses costeros de España. Emplearon 55 escenas emparejadas con *Chl-a in-situ* tomadas entre los años 2002-2008, desarrollaron tres diferentes modelos NN, uno entrenado con el set de datos completo, y los otros dos con los clústeres resultantes al aplicar el método *Fuzzy C-*

Mean. Los tres modelos obtuvieron un buen desempeño, pero la mejor predicción se dio con el cluster de valores Rrs más altos con resultados de 0.97 R2 en entrenamiento y 0.86 R2 en su validación.

- Métodos Kernel (Shawe-Taylor, Cristianini, *et al.,*2004), se dividen en: Máquinas de Soporte de Vectores (SVM, Support Vector Machines (Vapnik *et al.,* 1996)); Kernel Ridge Regression (KRR (Suykens *et al.,* 2000)) también conocido como Máquinas De Soporte de Vectores De Mínimos Cuadrados (Least Squares Support Vector Machines, LS-SVM); Máquinas de Vectores de Relevancia (Relevance Vector Machines, RVM (Samui *et al.,* 2008; Camps-Valls *et al.,*2006)); Regresión de Procesos Gaussianos (Gaussian Processes Regression, GPR (Rasmussen, Williams, *et al.,* 2006)). Chegoonian *et al.,* (2021) demostraron el potencial de SVR empleando datos Sentinel2 en un lago eutrófico de Canadá, argumentan que en cuerpos de agua pequeños donde no hay muchos pares coincidentes disponibles de *Chl-a in-situ* y reflectancia SVR puede obtener mejor precisión que los modelos empíricos empleados usualmente.

- Redes Bayesianas (Cooper & Herskovits, 1992). Werther *et al.,*(2022) realizaron un estudio sobre el aprendizaje automático bayesiano para la estimación y visualización de *Chl-a* empleando productos OLCI y MSI para lagos mesotróficos y oligotróficos pequeños y grandes, desarrollaron el algoritmo Redes Neuronales Probabilísticas Bayesianas (Bayesian probabilistic neural networks, BNN), concluyendo que el algoritmo BNN funcionan de manera similar para las observaciones de MSI y OLCI lo que sugiere su potencial para entregar productos de *Chl-a* de alta calidad con múltiples misiones.

El objetivo de investigación en este capítulo es generar una base de datos con mediciones de concentración de *Chl-a in-situ* y valores de Rrs OLCI en SAMAO para hacer el entrenamiento y validación de un conjunto de algoritmos de regresión de aprendizaje de máquina implementados en la herramienta ARTMO (Verrelst *et al.,*2012; Rivera-Caicedo *et al.,*2014). Como parte del marco metodológico se implementó la clasificación de los perfiles ópticos de SAMAO usando clasificación no supervisada (Moore *et al.,*2014) con el propósito de mejorar la precisión en la estimación de *Chl-a*.

Metodología

Generación de base de datos

Las bases de datos para teledetección son una fuente importante de información que provee mediciones de variables ambientales como la concentración de *Chl-a* y de reflectancia obtenida con sensores remotos. La concentración de *Chl-a* presente en las algas acuáticas es una medida biológica importante que se usa comúnmente para evaluar la biomasa total de las algas presentes en los cuerpos de agua (Jamie Bartram & Ballance, 1996).

Clorofila *in-situ*

Se realizaron siete muestreos de agua en el lago durante el año 2020 (tabla *1*). Se diseñó una red de muestreo con estaciones abarcando orillas y centro del lago (figura *5*). La toma de muestras se realizó en la parte superficial de la columna de agua usando contenedores opacos con tapa y refrigerándolas inmediatamente para evitar la degradación de las moléculas de *Chl-a*. Las muestras se trasladaron al laboratorio de Ciencias Ambientales en el Centro Nayarita de Innovación y Transferencia de Tecnología de la Universidad Autónoma de Nayarit, donde se

realizó la separación de las células del agua mediante filtración, y posteriormente la extracción y medición de *Chl-a* por el método fluorométrico.

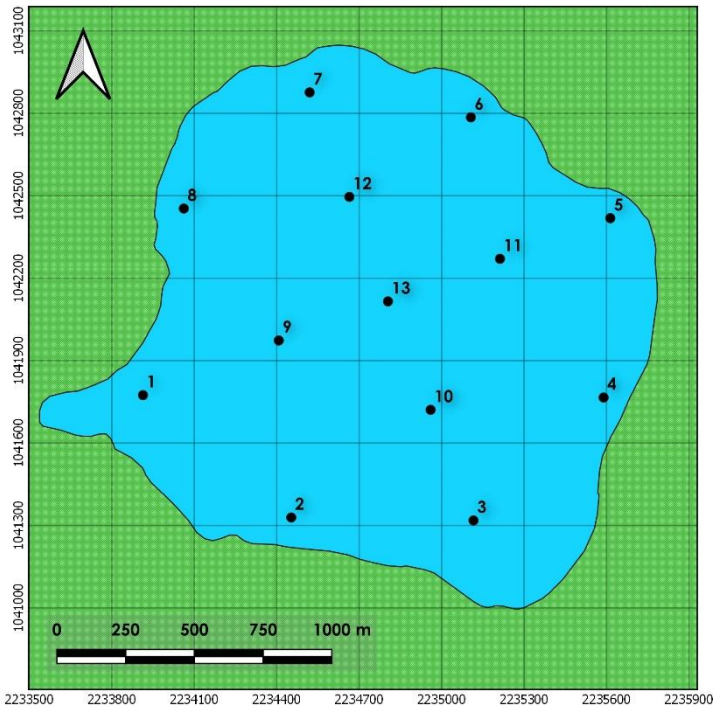


Figura 5 . *Ubicación espacial de red de estaciones de muestreo.*

Tabla 1. *Fechas de las campañas de muestreo.*

No.	Día	Mes	No. de muestras
M1	17	Enero	13
M2	28	Febrero	13
M3	19	Marzo	11
M4	3	Mayo	13
M5	29	Mayo	13
M6	02	Septiembre	13
M7	29	Septiembre	13

El contenido de *Chl-a* en muestras se usa comúnmente como un indicativo de la biomasa por fitoplancton presente en los cuerpos de agua. La medida de concentración de *Chl-a* se logra extrayendo el pigmento de las células con un solvente orgánico y relacionando sus unidades de fluorescencia relativa con concentraciones de *Chl-a* en $\mu g/L$.

i. Se utilizaron filtros de fibra de vidrio GF/F de 25 mm de diámetro (marca $Whatman^{TM}$) y se registraron los volúmenes iniciales de agua. Los filtros se colocan doblados por la mitad con el contenido hacia dentro en empaques individuales marcados y se almacenan a -20°C en la oscuridad.

ii. Para la extracción se empleó acetona al 90% como solvente y frascos opacos de vidrio con tapa. Después de descongelados, introducimos un filtro a cada frasco y agregamos 10 ml de acetona individualmente. Los frascos con el solvente y el filtro se almacenaron por 24 horas en un refrigerador a 4°C.

iii. Una vez pasado este tiempo en refrigeración, se decanta el sobrenadante de los frascos a tubos de ensayo con tapa y se ponen a centrifugar a 4500 revoluciones por minuto, a 20°C por 10 minutos.

iv. Para realizar la lectura utilizamos el equipo Trilogy Fluorometer modelo 7200-000 empleando el módulo para *Chl-a* no acidificada. Tomamos dos mililitros de la solución centrifugada desde la parte superior de los tubos de ensayo, los colocamos en las celdas de vidrio del equipo de fluorometría, ingresamos en la pantalla del equipo el volumen filtrado y el volumen del solvente para realizar la lectura. Previo a ésta, se llevó a cabo la calibración del equipo Trilogy mediante curvas de calibración realizadas con un estándar de 1 mg de *Chl-a* de *Anacystis nidulans* (C 6144 de Sigma Aldrich, Inc.), lo anterior para

convertir las unidades de fluorescencia relativa del equipo a unidades de concentración en $\mu g/L$.

Datos satelitales

COPERNICUS es el Programa de Observación de la Tierra la Agencia Espacial de la Unión Europea (ESA, por sus siglas en inglés). Este programa consiste de un conjunto de satélites (las familias Sentinel, figura *6*) y misiones contribuyentes (satélites comerciales y públicos existentes). COPERNICUS proporciona, a través de su constelación de satélites, un sistema unificado mediante el cual se obtienen grandes cantidades de datos de reflectividad casi en tiempo real. Actualmente tres constelaciones de dos satélites están en órbita, más un satélite adicional: Sentinel-1A (2014), Sentinel-1B (2016), Sentinel-2A (2015), Sentinel-2B (2017), Sentinel-3A (2016), Sentinel-3B (2018), Sentinel-5 (Precursor) (2017). Como parte del "Programa Europeo de Observación y Seguimiento de la Tierra", COPERNICUS estima colocar en órbita una constelación de casi 20 satélites más antes del 2030.

El sensor utilizado en este trabajo es OLCI, a bordo de la misión Sentinel-3 (S3). La resolución espacial de OLCI es de aproximadamente 300 m, mientras que su capacidad espectral va desde el visible hasta el infrarrojo cercano (de $\lambda 400nm$ a $\lambda 1200nm$) con sus con 21 bandas espectrales que realizan lecturas en forma simultánea. En la tabla *2* se muestra la información y el propósito general de cada banda espectral (información obtenida a través de la ESA, *https://sentinels.copernicus.eu*).

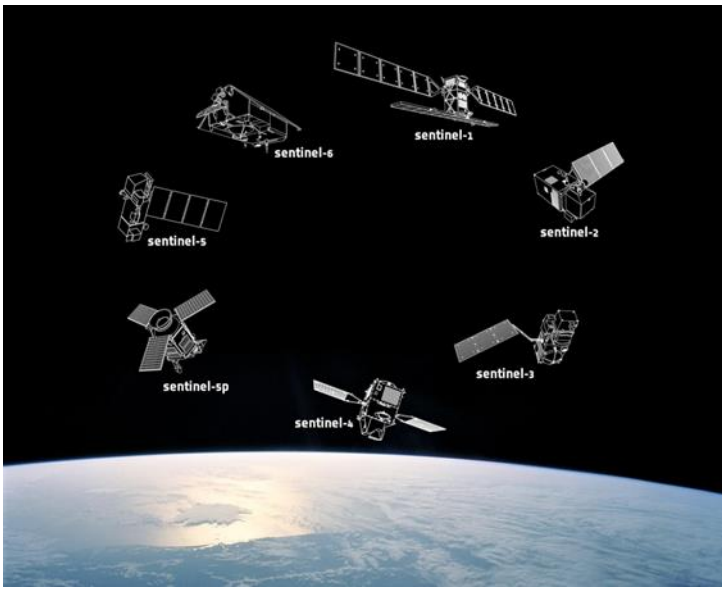


Figura 6. Misiones SENTINEL del programa COPERNICUS (www.esa.int)

SENTINEL3 incorporó diversas mejoras en sus resoluciones espectrales y temporales en comparación con otros sensores ópticos de observación terrestre. Con estas capacidades mejoradas, existe una gran motivación para desarrollar productos OLCI adecuadamente calibrados y validados para monitorear los riesgos de eutrofización y floraciones de algas tóxicas que, generalmente presentan objetivos desafiantes para la detección remota, pero tienen impactos económicos y sociales significativos (Kravitz *et al.,*2020). Así, al tiempo que nuevos sensores son diseñados y enviados al espacio, hay una demanda por mejorar las estrategias de recuperación operacional de los parámetros biofísicos relevantes para entender dinámicas ambientales locales y globales (J. Verrelst *et al.,*2011).

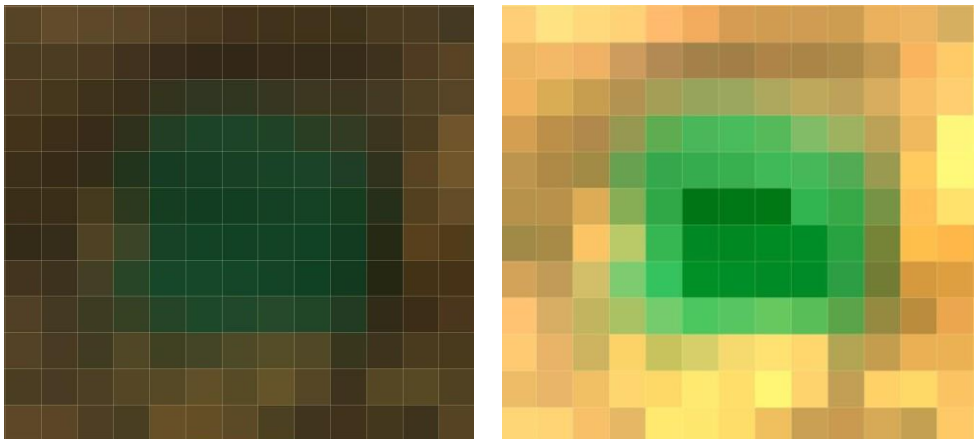
Tabla 2 Bandas espectrales *de OLCI.*

Banda	λ centro (nm)	ancho (nm)	Función
Oa01	400	15	Corrección de aerosoles (aerosol corr.)
Oa02	412.5	10	Sustancia amarilla (CDOM) y turbidez
Oa03	442.5	10	Máximo de abs. de *Chl-a*, biogeoquímica y vegetación
Oa04	490	10	*Chl-a* alta
Oa05	510	10	*Chl-a*, turbidez, sedimentos y mareas rojas
Oa06	560	10	Mínimo abs. *Chl-a*
Oa07	620	10	Descarga de sedimentos
Oa08	665	10	Segundo máximo de abs. de *Chl-a* y CDOM
Oa09	673.75	7.5	Fluorescencia
Oa10	681.25	7.5	Pico de fluorescencia de *Chl-a*, borde rojo
Oa11	708.75	10	Línea base de fluorescencia, transición del borde rojo
Oa12	753.75	7.5	O2 absorción/nubes, vegetación
Oa13	761.25	2.5	O2 banda de absorción/corrección de aerosoles
Oa14	764.375	3.75	Corrección atmosférica (Atmos. corr)
Oa15	767.5	2.5	O2A presión de la cima de la nube, fluorescencia
Oa16	778.75	15	Atmos. corr./aerosol corr.
Oa17	865	20	Atmos. corr./aerosol corr., nubes, co-registro pixel
Oa18	885	10	Banda de referencia de abs. del vapor de agua
Oa19	900	10	Abs. del vapor de agua/vegetación (máx. reflectancia)
Oa20	940	20	Abs. de vapor de agua, Atmos. corr./aerosol corr.
Oa21	1 020	40	Atmos. corr./aerosol corr.

Procesamiento de escenas

Desde la plataforma Copernicus Open Access Hub (https://scihub.copernicus.eu/) se obtuvieron los productos *'OL_1_EFR - Full Resolution TOA Reflectance'* disponibles para las fechas de las campañas de muestreo en campo. Éstos se descargaron con un nivel de procesamiento 1B, el cual corresponde a datos con corrección geométrica, calibración espectral y calibración radiométrica de la radiancia medida en el techo de la atmósfera (radiancia TOA, por las siglas en inglés de Top Of the Atmosphere). Para convertir los valores TOA a reflectividad de la superficie terrestre (Rrs) se realizó la corrección atmosférica empleando el algoritmo de redes neuronales C2RCC (Case 2 Regional Coast Colour) (Brockmann *et al.,*2016). En la figura 7 se muestra el recorte de SAMAO en el antes y después de aplicar la corrección.

C2RCC es un software para procesar datos de OLI, MERIS, MODIS, SeaWiFS, MSI y OLCI, que está disponible en la caja de herramientas del software de la Agencia Espacial Europea: Sentinels Application Platform (SNAP). Éste es empleado para generar productos de aguas Caso II. C2RCC se basa en una gran cantidad de datos de simulaciones de reflectancia saliente del agua relacionadas a valores de radiancias TOA (Brockmann *et al.,*2016) y es una modificación del Caso II Regional Processor (C2R) de (Doerffer & Schiller, 2007) hecha por el proyecto CoastColour (www.coastcolour.org).



A) Recorte nivel 1B *B) Recorte con la corrección atmosférica*

Figura 7. Recorte Sentinel-3A antes y después de la corrección atmosférica.

Extracción y caracterización de firmas espectrales

Para la extracción de las firmas espectrales del lago se utilizó el programa SNAP (Sentinel Application Platform) junto a la Sentinel-3 Toolbox. Con el fin de realzar los valores bajos medidos en las bandas a partir de los 650 nm, se aplicó una transformación logarítmica a los datos. Se analizaron las características espectrales de SAMAO mediante medidas de tendencia central (media, mediana, moda) y las medidas dispersión (rango, varianza, desviación típica). Se analizaron las topologías de las firmas espectrales que se presentaron en SAMAO durante el periodo de estudio. Estas topologías son firmas espectrales tipo que se han sugerido como mecanismos para delinear diferencias en el agua en función de sus propiedades ópticas (Spyrakos *et al.,*2018), la clasificación de los perfiles se realizó con el método no supervisado *K-means*. Debido a la sensibilidad de *K-means* a la ubicación inicial de los centroides (Jain 2010) se evaluaron de manera aleatoria 100 posiciones de éstos y se promediaron los resultados.

Para encontrar el numero óptimo de grupos se usó el Coeficiente de Silhouette (Starczewski & Krzyżak 2015). Este coeficiente mide la distancia de un punto x a los demás puntos de su misma clase como medida de cohesión $\big(a(x)\big)$ y la distancia media de x a los puntos de otras clases como medida de separación $\big(b(x)\big)$. Los valores de Silhouette van de -1 a +1, donde valores altos indican que el objeto está bien emparejado con su propio grupo y mal emparejado con el resto de grupos. Viene definido por la ecuación:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

$$SC = \frac{1}{N}\sum_{i=1}^{N} s\,(x)$$

Donde el valor de $s(x)$ puede variar entre -1 y 1 y SC es el coeficiente de Silhouette general y N el número de elementos clasificados.

Algoritmos de estimación de *Chl-a*

Se evaluaron cinco algoritmos regresión de aprendizaje automático o *MLRA* para estimar la concentración de *Chl-a*. La Tabla *5* muestra los algoritmos evaluados, cada uno de ellos cuenta con parámetros que pueden ser ajustados a fin de mejorar el rendimiento y la precisión de cada uno. La columna uno describe el nombre del algoritmo, en la columna dos se detalla la función utilizada por MATLAB para emplearlo y se detalla el hiperparámetro y los rangos que fueron evaluados, la columna 3 muestra la referencia bibliográfica.

Estos algoritmos están implementados en la herramienta *Simple-R* desarrollada por el Laboratorio de Procesamiento de Imágenes (Image Processing Laboratory- IPL), de la Universidad del Valencia (Image Processing Lab (IPL) 2016). Su interfaz gráfica de usuario se

ha implementado en el programa ARTMO (Operador Automatizado de Modelos de Transferencia Radiativa), éste es un software científico para procesamiento de datos satelitales y, fue desarrollado en MATLAB por la Universidad de Valencia (Rivera-Caicedo *et al.,*2014).

Tabla 3 . Algoritmos de aprendizaje de máquina para estimar concentración de Chl-a

Algoritmos de regresión	Función MATLAB	Referencia
Mínimos Cuadrados Parciales	plsregress No. Comp. Principales [2 - 10]	De Jong (1993); Rosipal & Kr¨amer (2006)
Random Forest	fitensemble No. Arboles [5,10,15,25,50,75,100,200]	Breiman (2001)
Redes Neuronales	fitnet No. neuronas/capas ocultas [2,4,6,8,10,12,14,16,18,20]	Duc-Hung *et al.*, (2012)
Kernel Ridge Regression	simple-R Interno	Pelckmans *et al.*, (2002)
Gaussian Process Regression	simple-R Interno	Williams & Rasmussen (2006)

Mínimos cuadrados parciales (PLS)

Este algoritmo combina las técnicas de análisis de componentes principales y la regresión lineal múltiple. Evita la multicolinealidad en las variables independientes y permite analizar problemas donde el número de muestras es menor al número de variables dependientes (Valdéz Blanco 2010). Extrae un conjunto de factores latentes que explica en la mayor medida posible la covarianza entre las variables dependientes e independientes usando análisis de componentes principales (Bro & Smilde, 2014).

El modelo subyacente general de PLS multivariantes es:

$$X = TP^T + E$$
$$Y = UQ^T + F,$$

Donde X es una matriz $n \times m$. El número de muestras es n, y m es el número de variables independientes, Y es una matriz de $n \times p$ donde p es el número de variables dependientes; T y U son matrices $n \times l$ donde l es el número de componentes seleccionados al realizar el análisis de componentes principales o proyecciones ortogonales que maximizan la varianza; P y Q son, respectivamente, $m \times l$ y $p \times l$ matrices de cargo ortogonales; y las matrices E y F son los términos de error.

Uno de los algoritmos más implementados al momento de solucionar PLS es el algoritmo SIMPLS (De Jong 1993). Este permite calcular los factores o pesos de la regresión directamente como combinación lineal de las variables originales.

Bosques Aleatorios (Random Forest- RF)

Un árbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico, con una estructura de árbol jerárquica, que se crea a partir de la información entregada por las variables independientes y sus relaciones con las variables dependientes utilizando diferentes estrategias para segmentar la información y de este modo crear reglas (De Ville 2013). Existen algoritmos que crean estas reglas tanto para problemas de clasificación o regresión (ID3, C4.5, CART y CHAID) (Batra & Agrawal, 2018). Este es el núcleo de los Bosques aleatorios que son conjuntos de múltiples arboles de decisión que han sido entrenados con técnicas de ensamble de modelos que reducen el sobre ajuste de los modelos (Sagi & Rokach, 2018). Cada árbol del conjunto de bosque aleatorio es entrenado de manera paralela. Obtenidos los resultados de cada árbol, en los

problemas de regresión se promedian los resultados para determinar la solución que entrega el bosque aleatorio.

Redes Neuronales (NN)

Una red neuronal es una técnica de aprendizaje de máquina que analiza los datos, inspirada en las relaciones sinápticas de las neuronas que conforman el cerebro humano. En la técnica de aprendizaje y minimización del error en las estimaciones se utilizó el gradiente conjugado escalado (Scaled Conjugate Gradient - SCG). Este es un método interactivo para resolver sistemas de ecuaciones lineales cuyas matrices son simétricas y definidas positivas. A partir de las condiciones previas, el cálculo de la dirección para encontrar el mínimo de errores en la estimación se optimiza, por lo general la búsqueda de la regresión es lineal, pero este método se ha diseñado para ser condicionalmente más óptimo (Babani *et al.,* 2016) y cuenta con tasa de convergencia superlineal (Roodschild *et al.,*2019). La topología de la red neuronal se muestra en la figura *8*, corresponde a una red multicapa donde las capas de entrada son cada banda del sensor OLCI, con una capa oculta donde se evalúan por defecto de 2 a 10 neuronas, W y b representan los parámetros que definen las rectas de las neuronas en la red.

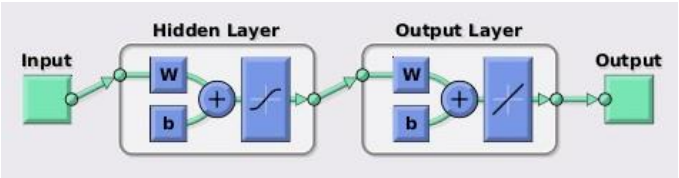


Figura 8. Topología tipo de una red neuronal usada por la función fitnet para la estimación de concentración de Chl-a.

Kernel Ridge Regression (KRR)

Esta técnica no paramétrica representa la unión de la regresión lineal "ridge" que vence el problema de la multicolinealidad de las variables independientes (Hoerl & Kennard 1970; Valle & Guerra, 2012) en problemas lineales con el *"truco Kernel"* que permite trabajar en problemas no lineares y no paramétricos. Este se basa en el concepto de espacio de Hilbert (Debnath & Mikusinski 2005), el cual es una generalización del concepto de espacio euclídeo y permite aplicar toda el álgebra euclidiana en espacios de alta dimensionalidad por medio de funciones Kernel. Los detalles de los diferentes Kernel utilizados pueden ser estudiados en Pompa (2018). La función Kernel utilizada aquí es *Radial Basis Function (RBF)* la cual puede ser representada así:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

Donde σ es un hiperparámetro libre y $\|x - x'\|$ es la diancia euclidiana entre los puntos x y x'. La implementación de *simple-R* parametriza la función Kernel con dos parámetros: σ y γ los cuales fueron evaluados para determinar la mejor regresión.

Gaussian Process Regression (GPR)

Este algoritmo no paramétrico se fundamenta en la regresión lineal bayesiana, cuya base radica en el teorema de Bayes (Efron, 2013). En dicho teorema se avalúan los parámetros del modelo no como estimaciones puntuales sino como distribuciones de probabilidad (Koduvely, 2015). De las distribuciones de probabilidad más frecuentes en la naturaleza es la gaussiana (Pértegas & Pita, 2001). A partir de la kernelización de esta regresión lineal tenemos la regresión

por procesos gaussianos. Los detalles de los diferentes Kernel utilizados pueden ser estudiados en Pompa (2018).

Algoritmo OC_4

El algoritmo de referencia para contrastar la hipótesis del presente estudio es el OC_4 propuesto por O'Reilly (2019), este algoritmo es el empleado por el *Ocean Biology Processing Group (OBPG)* de NASA para la generación de productos L2 y L3 del océano y CAC. Su ecuación es un polinomio de grado 4 como se muestra en la siguiente ecuación:

$$log_{10}(Chla) = a_0 + \sum_{i=1}^{4} a_i \left(log_{10} \left(\frac{R_{rs}(\lambda_{blue})}{R_{rs}(\lambda_{green})} \right) \right)^i$$

Donde, $R_{rs}(\lambda_{blue})$ es el máximo valor de reflectividad de las bandas entre las bandas $443\ nm$, $490\ nm$ y $510\ nm$ que corresponde a la región del azul del espectro electromagnético y $R_{rs}(\lambda_{green})$ corresponde a la banda $560\ nm$, en la región del verde del espectro electromagnético y $a_0, a_{1..4}$ corresponde a los coeficientes de polinomio.

Experimentos

Se generaron modelos de acuerdo a los resultados de la clasificación por la respuesta espectral obtenidos en el capítulo 1. Para esto se realizaron 8 experimentos (la tabla *6* para evaluar el desempeño de los MLRA en la estimación de la concentración de *Chl-a*, en los experimentos se evaluaron 3 estrategias que se enlistan a continuación:

- Transformación a escala logarítmica de los datos de reflectividad;
- Transformación a escala logarítmica de los datos de concentración de *Chl-a*; y

- Transformación logarítmica de datos Rrs y *Chl-a*

Tabla 4. Tratamientos para la evaluación de los parámetros

Experi-mento	Código	Descripción
T1	R_{rs} − *Chl-a* − 0C	Datos sin ninguna transformación ni clasificación supervisada
T2	R_{rs} − *Chl-a* − 2C	Datos sin ninguna transformación con clasificación supervisada de 2 clases
T3	$\log_{10}(R_{rs})$− *Chl-a* −0C	Datos con transformación logarítmica en los valores de la reflectancia y sin clasificación supervisada
T4	$\log_{10}(R_{rs})$− *Chl-a* −2C	Datos con transformación logarítmica en los valores de la reflectancia y con clasificación supervisada de 2 clases
T5	R_{rs} − $\log_{10}$ (*Chl-a*) −0C	Datos con transformación logarítmica en los valores de la *Chl-a* y sin clasificación supervisada
T6	R_{rs} − $\log_{10}$ (*Chl-a*) −2C	Datos con transformación logarítmica en los valores de la *Chlo-a* y con clasificación supervisada de 2 clases
T7	$\log_{10}(R_{rs})$− $\log_{10}$ (*Chlo-a*) − 0C	Datos con transformación logarítmica en los valores de reflectancia y la *Chl-a* y sin clasificación supervisada
T8	$\log_{10}(R_{rs})$− $\log_{10}$ (*Chl-a*) − 2C	Datos con transformación logarítmica en los valores de reflectancia y la *Chl-a* y con clasificación supervisada de 2 clases

Validación de modelos

Para la validación de la precisión de los algoritmos se realizó el método de dejar uno fuera o Leave-one-out cross-validation (LOOCV) (Efron, 1982), este método de validación es comúnmente usado cuando se cuenta con una base de datos de pocas muestras. El estadístico que se empleó para la evaluación del rendimiento del algoritmo es el Error Cuadrático Medio o $RMSE$:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}$$

Donde, $\hat{y}_t$ son los valores de la concentracion de *Chl-a* estimada por el modelo de regresión, y_t es la concentración de *Chl-a* medida en campo por técnicas de fluorometría y T son el numero de muestras. El modelo más preciso será aquel con menor $RMSE$.

Resultados

Concentración de clorofila *in-situ*

Se determinó la concentración de *Chl-a* para las siete campañas de campo realizadas según el cronograma que se muestra en la tabla *1*. El gráfico de la figura *10* muestra los valores de la concentración de *Chl-a*, donde en el eje $'y'$ están los valores determinados por fluorometría, en el eje $'x'$ las trece estaciones de los muestreos mientras que los colores las barras identifican cada campaña de campo.

El análisis de la distribución temporal de los valores medidos en la campaña 2020 se muestran en la figura *9*. El gráfico de caja muestra la dispersión de los valores, los percentiles 25 y 75 de los datos y los limites superiores e inferiores identifican los valores atípicos en las series. La distribución espacial de los valores presenta sesgos positivos en cada muestreo. La mayor concentración de *Chl-a* se obtuvo en los muestreos realizados en marzo (M3) y mayo (M5), y la menor en febrero (M2). La mediana de los datos en todos los muestreos está por debajo de 11.0 $\mu g/L$, sin embargo, hubo algunas estaciones que alcanzaron valores por arriba de los 40 $\mu g/L$. Las concentraciones de *Chl-a* en la mayoría de los muestreos (excepto en el M2) tuvieron alta dispersión en los valores más altos.

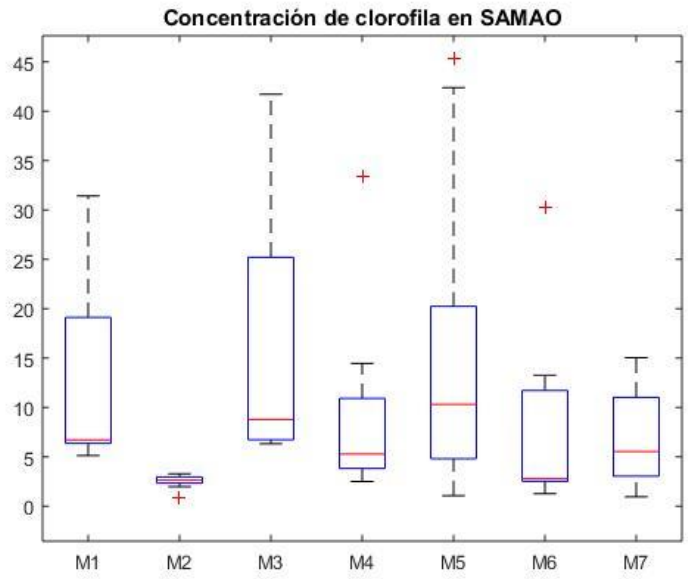


Figura 9. Distribución de los valores de Chl-a agrupados por campaña de campo.

Como se muestra en la tabla 5, el valor máximo de *Chl-a* se registró en la estación 2 con un valor de 45.3 $\mu g/L$, seguido de la estación 1 con 42.4 $\mu g/L$ ambos en el muestreo M5. Los valores más bajos se registraron en las estaciones 2 y 6 con valores de 0.9 $\mu g/L$ y 0.96 $\mu g/L$ en los muestreos M2 y M7 respectivamente. En relación al valor promedio medido en toda la campaña 2020, la estación 2 presento el mayor promedio con un valor de 21.65 $\mu g/L$ y la estación 11 el menor valor con 4.48 $\mu g/L$. La estación con mayor variabilidad en sus medidas fue la estación 2 con una desviación estándar de 18.04 $\mu g/L$ y la estación 11 fue la más homogénea con un valor de 2.44 $\mu g/L$.

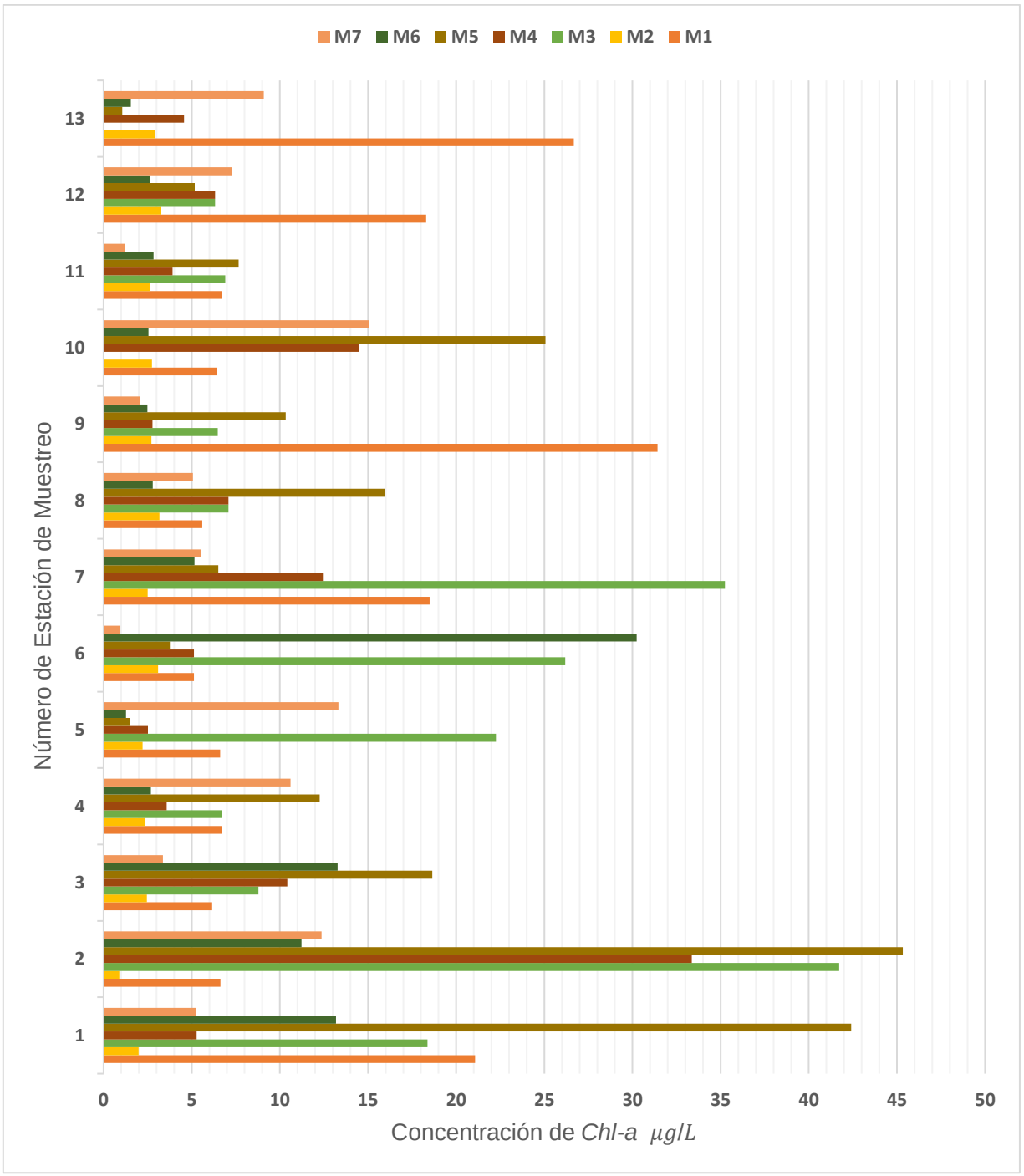


Figura 10. Concentración total de Chl-a por estación de muestreo.

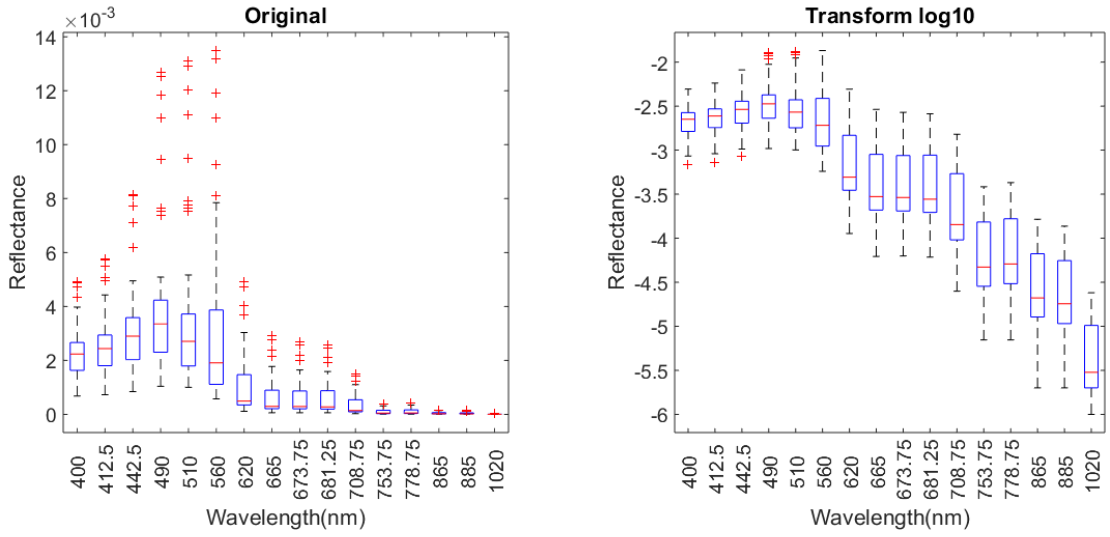
Tabla 5. Estadísticos descriptivos de concentración de Chl-a en relación a las estaciones de muestreo.

Id. Est.	Max.	Min.	Media	Desv. Std	Media-na	Rango Inter-cuartil
1	42.40	1.99	15.37	13.90	13.19	15.13
2	45.33	0.90	21.65	18.00	12.36	31.86
3	18.65	2.45	9.02	5.70	8.79	8.50
4	12.26	2.37	6.42	3.90	6.69	6.72
5	22.26	1.27	7.10	8.00	2.51	9.98
6	30.23	0.96	10.64	12.10	5.12	17.66
7	35.24	2.50	12.27	11.50	6.50	11.73
8	15.96	2.79	6.68	4.40	5.60	3.45
9	31.42	2.05	8.32	10.60	2.78	6.82
10	25.07	2.54	10.46	8.20	1.45	12.30
11	7.66	1.21	4.48	2.40	3.92	4.17
12	18.29	2.65	7.17	5.70	6.33	3.30
13	26.67	1.06	7.65	9.80	3.76	7.53

Caracterización de las firmas espectrales en SAMAO

La Figura *11* muestra la caracterización espectral del conjunto de 7 escenas OLCI asociadas a la toma de muestras *in-situ*. Se compararon los valores Rrs sin transformación (Rrs_{normal}) y transformados a escala logarítmica en base 10(Rrs_{log}), la dispersión de valores de las 16 bandas OLCI se muestran mediante diagramas de caja. En *11a* se observa la variabilidad en la reflectividad Rrs_{normal}, dónde la mayor variabilidad se presenta en la región del visible entre 400 y $560\ nm$, dentro de ésta los picos de reflectancia se dan en el espectro del verde y la banda con mayor variabilidad es la centrada en los $560\ nm$ la cual que coincide con el mínimo de absorción de *Chl-a*. A partir de esta banda, la variabilidad de los datos disminuye presentando su mínimo en la zona del infrarrojo. La figura *11b* muestra la distribución de $Rrs_{\log}$, ésta presenta una mayor dispersión de los datos de reflectancia en todas sus bandas, sin embargo, su comportamiento es inverso a Rrs_{normal} pues los datos de la región visible del espectro electromagnético tienen menor variabilidad, y en general presenta menor cantidad de datos atípicos en todas las bandas.

Los resultados de la figura 12 muestran la comparación de la capacidad discriminante analizada con el Coeficiente Silhouette (SC), donde la mejor separación de clase se obtiene cuando se seleccionan 2 clases, obteniendo un SC de 0.91 y 0.76 para Rrs_{normal} y Rrs_{log} respectivamente. La tabla *6* muestra los estadísticos de tendencia central y medidas de dispersión para la base de datos Rrs_{log} divididos en más de dos clases. En promedio la mejor separabilidad se da con 8 clases (SC 0.6712), sin embargo, al revisar los estadísticos los mejores resultados se dan con 7 clases con un máximo de SC 0.7236 y una moda de SC 0.7196.



a) Rrs_{normal} b) Rrs_{log}

Figura 11. Caracterización de firmas espectrales. A) Reflectividad (sr-1); B) sr-1 con transformación logarítmica.

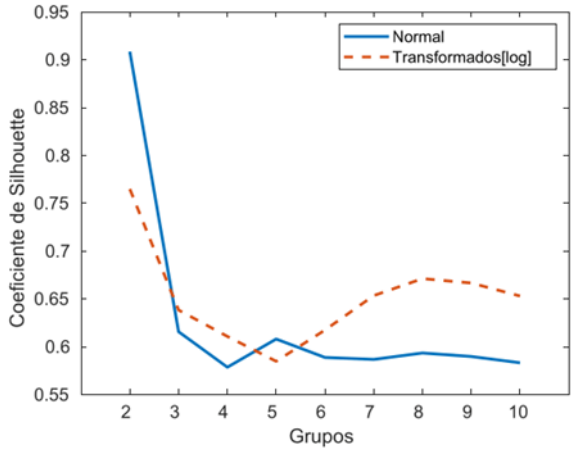
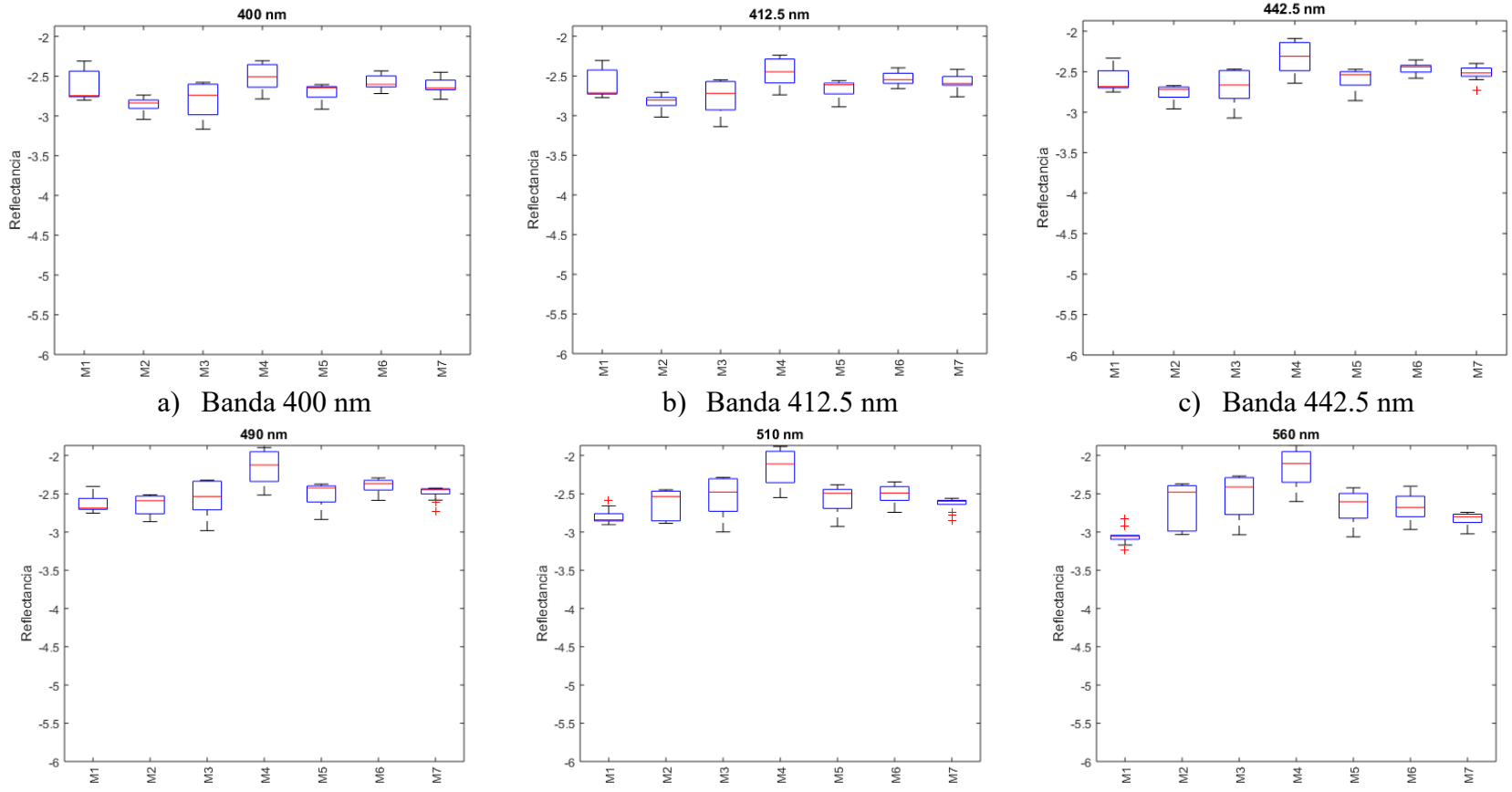


Figura 12. Comparación de la capacidad discriminante analizada con el Coeficiente Silhouette (SC) por cada uno de los grupos determinados con K-means para las bases de datos Rrs $_{nosmal}$ y Rrs $_{log}$. En el eje 'y' el valor de SC, en el eje 'x' el número de clases analizado.

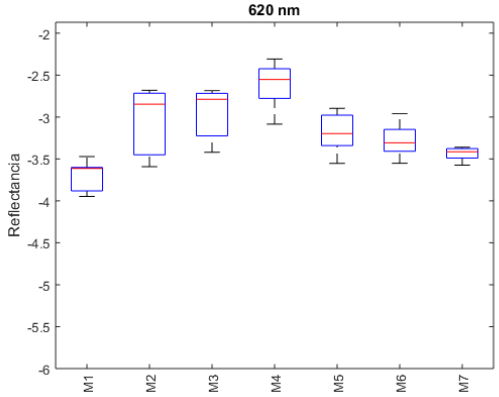
La figura *13* muestra la variabilidad temporal en cada una de las campañas realizadas en 2020 de cada una de las bandas del sensor OLCI con los valores de reflectancia Rrs_{log10} por medio de diagramas de cajas. Las figuras *13a*, *13b* y *13c* se muestra que corresponde la región del azul en el espectro electromagnético, presentan una distribución bimodal sus dos mínimos locales en M2-M3, correspondientes a las fechas 28 de febrero y 19 de marzo, y tres máximos locales en los muestreos M1, M4 y M6 realizados en las fechas 17 de enero, 03 de mayo y 29 de mayo respectivamente. La región del verde (figuras: *13d, 13e* y *13f*) presenta una distribución simétrica con el máximo en el M4 del 03 de mayo. La región del rojo (figuras: *13g*,*13h*, *13i* y *13j*) y el NIR (figuras: *13k*, *13l 13m, 13n*) tienden a presentar una distribución con sesgo positivo con una mayor variabilidad en los meses de febrero y marzo (M2-M3), y la menor variabilidad en el M7 del 29 de septiembre.



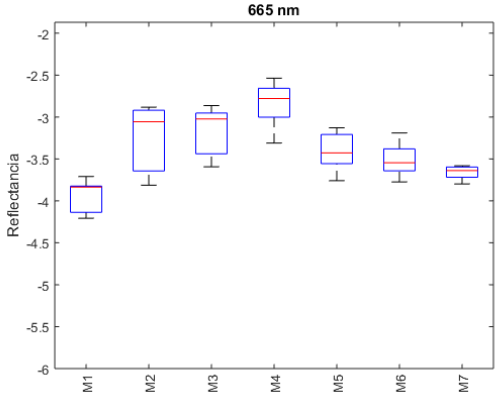
a) Banda 400 nm b) Banda 412.5 nm c) Banda 442.5 nm

d) Banda 490 nm e) Banda 510 nm f) Banda 560 nm

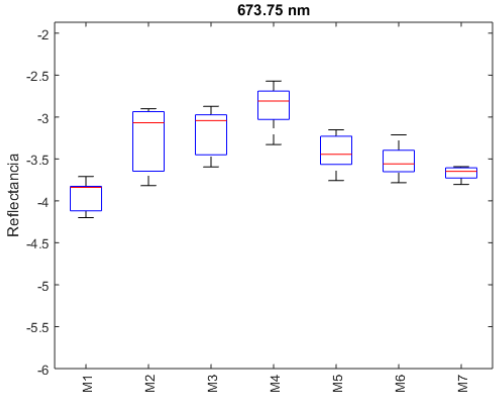
Figura 13. Caracterización de datos Rrs_{log10} OLCI asociados a los 7 muestreos en campo.



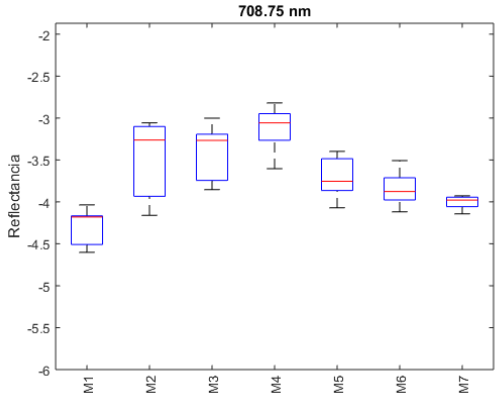
g) Banda 620 nm



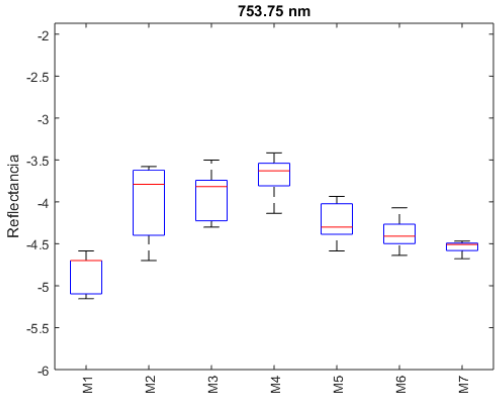
h) Banda 665 nm



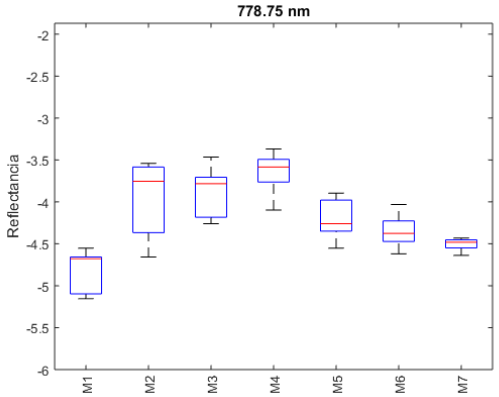
i) Banda 673 nm



j) Banda 708.75 nm

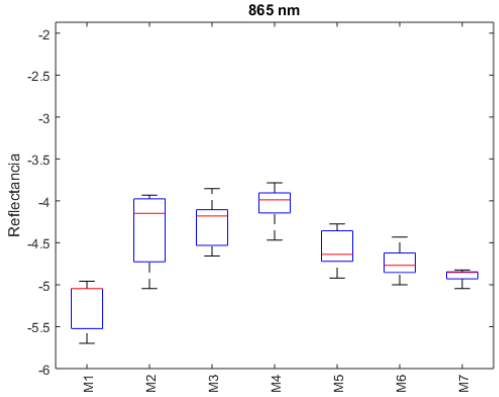


k) Banda 753.75 nm

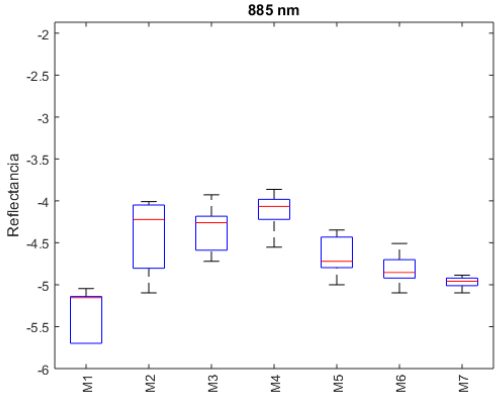


l) Banda 778.75 nm

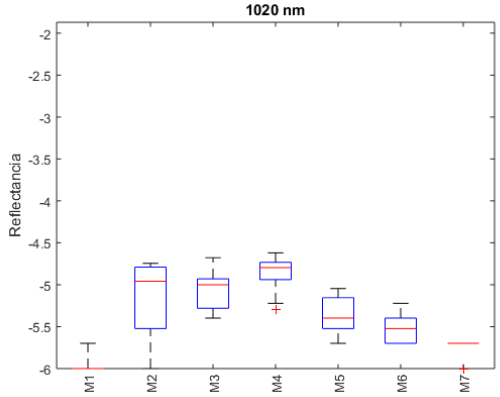
Figura 13. Caracterización de datos Rrs_{log10} OLCI asociados a los 7 muestreos en campo (Cont.).



m) Banda 865 nm



n) Banda 885 nm



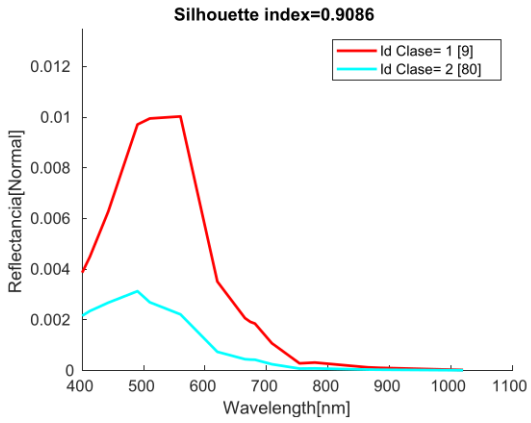
ñ) Banda 1020 nm

Figura 13. Caracterización de datos Rrs_{log10} OLCI asociados a los 7 muestreos en campo (Cont.).

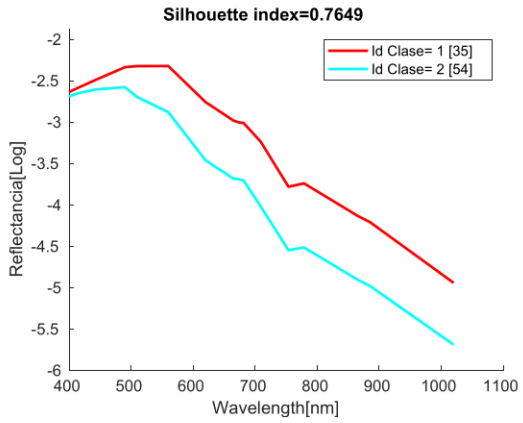
Tabla 6. Estadísticas del Coeficiente Silhouette para la base de datos de reflectivivdasd en escala logarítmica (Rrslog)

Estadístico	Número de clases								
	2	3	4	5	6	7	8	9	10
Promedio	0.76	0.63	0.61	0.58	0.62	0.65	**0.67**	0.66	0.65
Mediana	0.76	0.62	0.59	0.58	0.62	0.65	0.69	0.67	0.65
Moda	0.76	0.60	0.66	0.60	0.62	**0.72**	0.70	0.70	0.67
Mínimo	0.76	0.49	0.47	0.41	0.48	0.48	0.47	0.51	0.42
Máximo	0.76	0.71	0.66	0.65	0.68	**0.72**	0.71	0.72	0.71
Varianza	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Desviación Típica	0.00	0.05	0.05	0.04	0.05	0.07	0.05	0.04	0.04

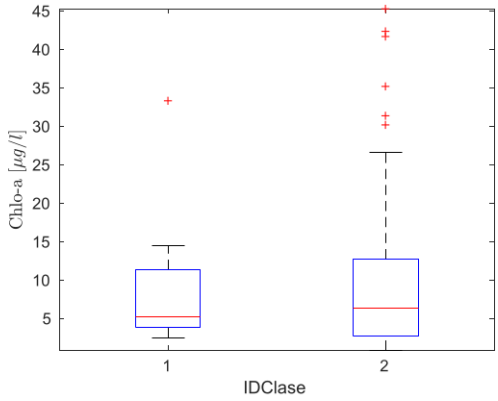
La Figura *14* muestra los perfiles promedios de las bases de datos Rrs_{normal} y Rrs_{log} agrupadas para dos clases y la distribución de la concentración de *Chl-a* en cada una de ellas. Aunque los resultados de SC para dos clases fue mejor en la base de datos Rrs_{normal}, en la figura *4a* observamos el número de muestras que conforman la clase 1 y 2, y por el número de perfiles que pertenecen a cada una de éstas (9 y 80 respectivamente) salta a la vista que la transformación logarítmica ayuda a eliminar el sesgo en la clasificación de los datos y a generar clases más balanceadas como se muestra en la figura *14b*.



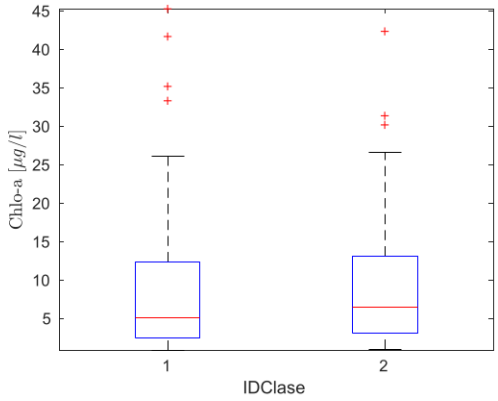
a) Perfiles espectrales para Rrs_{normal}



b) Perfiles espectrales para Rrs_{log}



c) Distribución de la *Chl-a* (Rrs_{normal})



d) Distribución de la *Chl-a* ($Rrslog$)

Figura 14. Perfiles y distribución de Chl-a en función de las clases definidas. a) Perfiles espectrales para Rrs_{normal}; b) Perfiles espectrales para Rrs_{log}, c) y d) gráficos de caja con la distribución de los datos en las clases.

Evaluación de algoritmos de regresión para *Chl-a*

Se presentan los resultados de la evaluación de 5 MLRAs con una BD conformada por 89 firmas espectrales de OLCI con su valor asociado de concentración de *Chl-a*, para lo cual se

desarrollaron 8 experimentos para encontrar la mejor estrategia de estimación de concentración de *Chl-a* en SAMAO.

Tabla 7. Desempeño (RMSE) de los MLRA's en función de cada experimento

Experi-mento	Código	PLS	RF	NN	KRR	GPR
T1	$Rr_s - Chlo - a - 0C$	**10.05**	10.39	**10.24**	11.60	**10.50**
T2	$Rr_s - Chlo - a - 2C$	10.90	10.47	10.79	11.39	10.91
T3	$log_{10}(Rr_s) - Chlo - a - 0C$	10.06	**10.37**	10.63	**10.72**	10.23
T4	$log_{10}(Rr_s) - Chlo - a - 2C$	10.48	10.49	12.19	12.56	10.51
T5	$Rr_s - \log_{10}(Chlo - a) - 0C$	10.05	10.39	10.24	11.60	10.51
T6	$Rr_s - \log_{10}(Chlo - a) - 2C$	10.90	10.47	10.79	11.39	10.91
T7	$log_{10}(Rr_s) - \log_{10}(Chlo - a) - 0C$	10.05	10.37	10.63	10.72	10.23
T8	$log_{10}(Rr_s) - \log_{10}(Chlo - a) - 2C$	10.48	10.49	12.19	12.56	10.51

La calibración del modelo de referencia OC_4 entregó los coeficientes: $a_0 = 0.8301$, $a_1 = -0.2386$, $a_2 = 0.0591$, $a_3 = 0.1944$ y $a_4 = 0.0685$, y obtuvo un $RMSE$ de 10.8. La tabla 7 resume los resultados obtenidos por cada uno de los 8 experimentos propuestos. El mejor resultado se obtuvo para la regresión por mínimos cuadrados parciales con 6 componentes principales con un valor de $RMSE = 10$. La diferencia entre los otros métodos estuvo en un rango entre el 3% al 7%. El modelo que presenta un comportamiento más homogéneo en todos los experimentos fue *Random Forest* con una desviación estándar entre los resultados de los

experimentos del 0.054. Mientras que las *Redes neuronales* presentaron la mayor diferencia con un valor std de 0.079.

Para valorar el desempeño de los algoritmos evaluados en la fase de entrenamiento, donde se genera el modelo para estimar concentración de *Chl-a,* se calculó el *RMSE* entre las muestras usadas en el entrenamiento y el valor estimasdo por el modelo. La tabla *8* muestra los resultados obtenidos. En esta evaluación es claro que los mejores modelos son los de tipo kernel: *Kernel Rigde* y *Guassian Processes*, con un valor del RMSE de 0.9308 y 1.1301 respectivamente. Aquí también se observa el impacto de la generación de modelos por cada tipo de respuesta óptica en SAMAO, ya que tres de los cinco modelos presentan sus mejores resultados con los experimentos donde se trabaja la base de datos con dos clases.

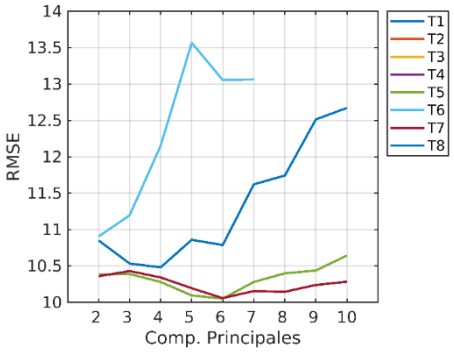
Tabla 8. Desempeño (RMSE) para los algoritmos en función de cada experimento en la fase de entrenamiento

Experi-mento	Código	PLS	RF	NN	KRR	GPR
T1	$Rr_s - Chlo - a - 0C$	9.3	**7.2**	7.4	3.3	**1.1**
T2	$Rr_s - Chlo - a - 2C$	8.6	7.4	7.8	**0.9**	3.1
T3	$log_{10}(Rr_s) - Chlo - a - 0C$	9.0	7.2	7.5	7.5	4.0
T4	$log_{10}(Rr_s) - Chlo - a - 2C$	**8.1**	7.6	**6.6**	3.3	5.3
T5	$Rr_s - \log_{10}(Chlo - a) - 0C$	9.0	7.2	7.4	3.3	1.1
T6	$Rr_s - \log_{10}(Chlo - a) - 2C$	8.6	7.4	7.8	0.9	3.1
T7	$log_{10}(Rr_s) - \log_{10}(Chlo - a) - 0C$	9.0	7.2	7.5	7.5	4.0
T8	$log_{10}(Rr_s) - \log_{10}(Chlo - a) - 2C$	8.1	7.6	6.6	3.3	5.3

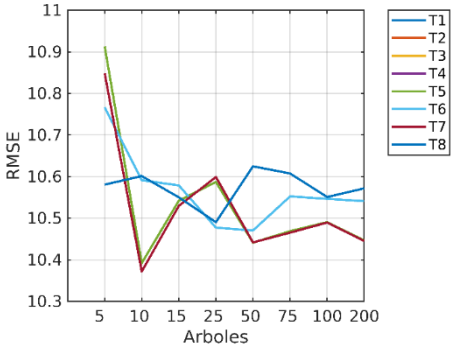
Evaluación de los hiperparámetros

La figura *15* muestra la evaluación de los hiperparámetros para los modelos PLS, RF y NN. En el eje y se muestra el RMSE calculado por cada modelo usando la validación LOOCV por su parte, el eje x muestra cada hiperparámetro en función de su respectivo modelo de estimación. La figura *15a* muestra el número de componentes principales evaluados en el modelo de mínimos cuadrados parciales, los mejores resultados se obtuvieron para los experimentos T5 y T7 con 6 componentes principales. En la figura *15b* vemos la evaluación para el algoritmo de Random Forest con diferente número de árboles donde, el número óptimo fue de 10 y, al igual que PLS, los mejores experimentos fueron los T5 y T7. Como se muestra en la figura *15c*, el mejor resultado en la evaluación del número de neuronas en la capa oculta de la red neuronal se obtuvo con 2 neuronas para el experimento T5 y 4 neuronas en la capa oculta con el experimento T7.

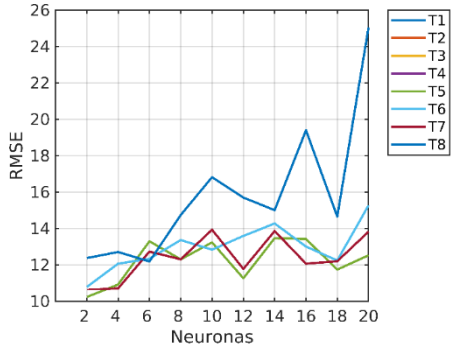
La figura *16* muestra los gráficos de dispersión en la concentración de *Chl-a* medida en campo en el eje x y la estimada por los modelos en el eje y. Se observa que los 5 algoritmos tendieron a subestimar los valores de *Chl-a* altos, y en mayor y menor grado sobreestimar los valores bajos medidos.



a) PLS

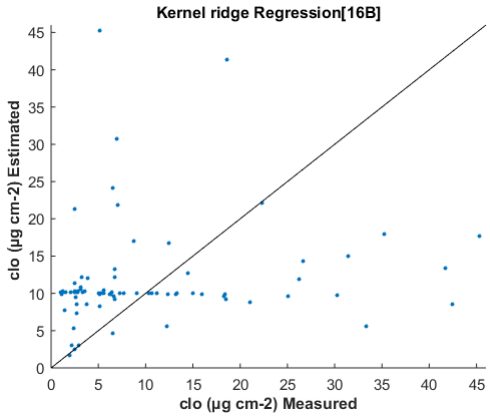


b) RF

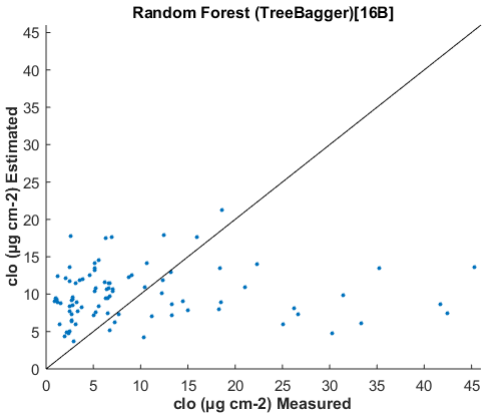


c) NN

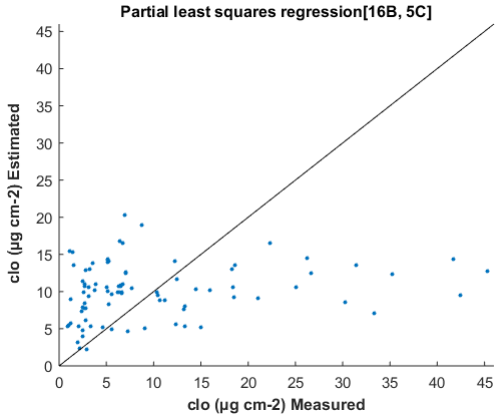
Figura 15. Evaluación de los hiperparámetros de los modelos a) Mínimos cuadrados parciales (PLS), b) Random Forest RF y c) Redes neuronales NN.



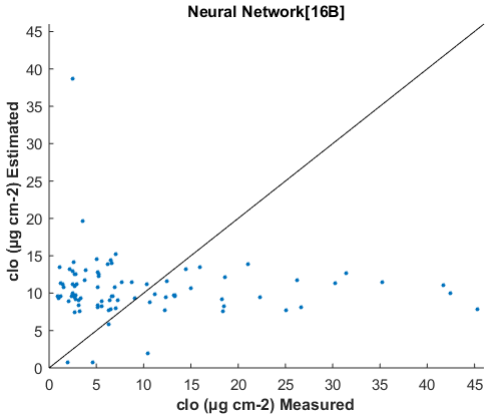
a) Kernel Ridge regression



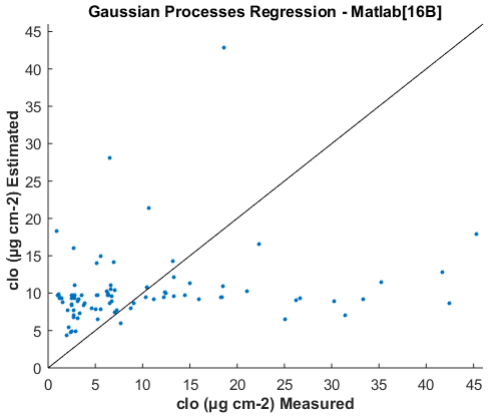
b) Random forest



c) Mínimos cuadrados parciales



d) Redes neuronales



e) Gaussian process

Figura 16. Dispersión entre Chl-a in-situ y Chl-a satelital.

Discusión

Se construyó una base de datos conformada por valores de concentración de *Chl-a* y valores de reflectividad medidos por el sensor OLCI para 93 muestras. En una comparación interanual encontramos que las concentraciones de *Chl-a* concuerdan con la dinámica de comportamiento reportado por Cortés (2018), quien midió mensualmente la concentración del pigmento en el lago durante el año 2015 reportando que el período de enero a abril fue el más productivo, mientras que en este estudio el período más productivo fue de enero a mayo, exceptuando el muestreo realizado en febrero que tuvo valores de 2.54 μg/L en promedio para los trece puntos de muestreo. El período de menor productividad es congruente con lo reportado por Cortés (2018) en 2015 dónde abarca los meses de mayo a diciembre. Al realizar una comparación con los resultados de este trabajo, en el mes de septiembre encontramos que las concentraciones del pigmento son en promedio diez veces superiores a lo reportado en 2015 en las 13 estaciones de muestreo. En este aspecto, Hernández *et al*. (2014) señala que las altas densidades celulares en la superficie de la columna de agua en lagos tropicales profundos durante primavera-verano pueden extenderse hasta otoño dependiendo de la cantidad de nutrientes disueltos en el agua. Esta cantidad de nutrientes está relacionada con la mezcla de la columna de agua, la cual a su se vincula a cambios regionales de temperatura. SAMAO presenta un régimen monomíctico cálido (Salas, 2017), esto quiere decir que su columna de agua se estratifica desde primavera tardía hasta principios de otoño (Lewis, 1983), mientras que en invierno la disminución de la temperatura de la capa superficial en la columna de agua provoca una mezcla que puede ser parcial o total en los inviernos más fríos, esto provoca un resurgimiento de los nutrientes depositados al fondo del lago en el hipolimnium. La precipitación también juega un papel crucial

en la disponibilidad de nutrientes (Méndez Reyes *et al.* 2018) especialmente en lagos-cráter como SAMAO. Ello se debe a que pues la cuenca de éste presenta escorrentía, la cual arrastra sedimentos, en este caso cargados de nitrógeno y fósforo desde los campos de cultivo donde se administran fertilizantes antes del periodo de lluvias.

Como lo indica Moore *et al.,* (2014), la clasificación de agua en Caso I y Caso II para CAC no es un sistema de clasificación adecuado, especialmente cuando se busca obtener valores de parámetros biofísicos mediante sensores remotos. Las propiedades ópticas de los CAC son particulares a cada uno de ellos, éstas están relacionadas con condiciones locales como: la cantidad y el tipo de sustancias disueltas en el agua, los efectos de adyacencia, dimensiones y profundidad del lago, así como el tipo de clima donde se ubica (Uudeberg *et al.,* 2019). En este sentido, el lago cráter de SAMAO, se caracteriza por tener cambios de coloración en una escala estacional como una respuesta óptica a variaciones en la biomasa y composición comunitaria del fitoplancton (Ochoa, 2018). Por lo que la preclasificación de la reflectancia permitirá generar un mejor algoritmo para la medición de la clorofila-a en SAMAO. Al respecto Shi *et al.,* (2013), Spyrakos *et al.,* (2018), Zhang *et al.,* (2019), Uudeberg *et al.,* (2019), Cui *et al.,* (2020) y Jiang *et al.,* (2020) han demostrado que la preclasificación de la reflectancia satelital en varios tipos de agua basada en las similitudes y diferencias de la forma de su espectro, es un esquema válido para reducir errores en la estimación de *Chl-a*, lo cual permite desarrollar algoritmos específicos a cada tipo de agua. En este estudio se observó que la mayor separabilidad de espectros de reflectividad se obtuvo con 2 clases. A pesar de lo anterior, éstas no guardan una relación directa con las concentraciones de *Chl-a* medidas en campo. Ello puede deberse al desfase de fechas en la toma de muestras con el paso del sensor (Palmer *et al.,*2015), a los desplazamientos en vertical

de las poblaciones de cianobacterias (Kutser *et al.,*2006) o al método de corrección atmosférica empleado (Toming *et al.,* 2017).

En tal sentido, el algoritmo empleado en este trabajo fue el C2RCC. Este algoritmo fue desarrollado para MERIS y adaptado para los datos de OLCI, sin embargo, en trabajos donde se le evaluó en comparación con otros algoritmos como ALTNNA (Uudeberg *et al.* 2019) y Alternative Atmospheric correction (AAC) (Ogashawara, 2019), no mostró una mejora significativa en relación a éstos. Por otro lado, Kravitz *et al*. (2020) encontró mejores resultados con el algoritmo 6SV1 (Kotchenova & Vermote, 2006) al evaluar los softwares iCOR, POLYMER y C2RCC, en relación con espectros medidos *in-situ* con radiómetro durante sus campañas de muestreo. Kravitz *et al*., (2020) también señalan que particularmente C2RCC solo debe ser aplicado a píxeles de agua pura, en ambientes más oligotróficos a mesotróficos. Así mismo encontró que, aunque C2RCC permitió observar la variabilidad espacial de los Florecimientos Algales, éste sobreestimó las muestras de menor concentración y subestimó el resto, concluyen que los resultados de C2RCC no tuvieron correlación con los datos medidos en campo. El principal problema puede ser la base de datos que alimenta el algoritmo, el número y la variación de perfiles espectrales que entrenan la red neuronal (Alcântara *et al.*, 2018). En este aspecto, en el año 2021 C2RCC fue actualizado incluyendo un nuevo procesador de corrección atmosférica en aguas ópticamente complejas: C2X-COMPLEX (C2XC) (Soriano-González *et al.*, 2022), el cual permitirá extender el potencial para mejorar las estimaciones de reflectancia a un rango más amplio de CACs.

El tratamiento de los datos de reflectancia con transformaciones en escala logaritmo en base 10 permite aprovechar toda la información espectral ofrecida por OLCI. El trabajo de Ayeni y

Adesalu (2018) demuestra cómo resulta conveniente transformar a escala logarítmica los valores de Rrs con el objeto de analizar su relación con la concentración de *Chl-a in-situ*. Para el entrenamiento de algoritmos de MLRA de estimación de *Chl-a,* se realizaron experimentos con la base de datos en los cuales se empleó la conversión de los datos a logaritmo base 10 y la clasificación en dos clases de acuerdo a sus perfiles espectrales. Sin embargo, la clasificación empleada no tuvo gran impacto, pues en la evaluación se obtuvieron mejores resultados con la base de datos completa. Como se esperaba, hubo una mejoría en la estimación con MLRAs respecto al algoritmo OC_4. La diferencia de su RMSE con el calculado para el MLRA de mejor rendimiento (mínimos cuadrados parciales) fue de 0.7238.

Conclusiones

El objetivo planteado para este capítulo fue alcanzado, se generó una base de datos de reflectividad OLCI y concentración de clorofila superficial *in-situ* medidas durante el año 2020. Sin embargo, durante el período de estudio, se pudieron realizar solo 7 de las 12 campañas de muestreo planeadas, de modo que se caracterizó parcialmente los tipos ópticos de agua que hay en SAMAO (Homogéneo, con Florecimiento y Turquesa). Los muestreos no pudieron ser realizados desde septiembre a diciembre por restricciones en el acceso al lago debido a la pandemia de COVID19.

Hay que destacar que contrario a lo esperado, los valores más bajos de biomasa se midieron en el muestreo de febrero, lo que nos indica que la dinámica de la biomasa en SAMAO cambia semanalmente. Lo anterior ya que, en observaciones con Sentinel 2 MSI con fechas del 17 y el 22 de febrero 2020 se percibió un florecimiento algal en la zona norte (17/02/2020) y noreste

(22/02/2020) del lago, mientras que en el muestreo *in-situ* que se realizó el 28 de febrero las concentraciones de *Chl-a* no superan los 3.5 $\mu g/L$.

Por otro lado, las mediciones de *Chl-a* presentaron una amplia variabilidad espacial en cada muestreo, lo que resalta que, a pesar de ser un CAC de pequeño tamaño, SAMAO no puede ser caracterizado con pocas muestras. Como señala Congalton (1991), cada punto de muestra recolectado es costoso y, por lo tanto, el tamaño de la muestra debe mantenerse al mínimo, sin embargo, es fundamental mantener una cantidad lo suficientemente grande de tamaño de la muestra para que cualquier análisis realizado sea estadísticamente válido. Las mediciones de biomasa en SAMAO demostraron que es un cuerpo de agua con amplio dinamismo espacial y temporal en sus niveles de concentración de *Chl-a*, las cuales varían de entre 0.9 hasta 45.4 µg/L, razones por las cuáles el período y número de toma de muestras en campo son componentes claves al realizar calibraciones de algoritmos para estimación de parámetros biofísicos. En estudios posteriores es necesario diseñar y realizar los muestreos en campo acorde a las necesidades específicas del área de estudio y las capacidades espaciales y temporales del sensor a fin de obtener muestras representativas.

La escala logarítmica en base diez permite observar mejor la distribución de las firmas espectrales, encontrar sus diferencias y separarlos en clases más adecuadas; en esta escala todas las bandas aportan información para clasificar. Al realizar un análisis de correlación de bandas con el método K-means para clasificar los perfiles espectrales de los datos OLCI. Los mejores resultados de SC se dieron al categorizar en 2 clases los datos Rrs_{normal} y $Rrs_{\log}$, sin embargo, esta clasificación no es suficiente para representar los estados tróficos del lago, en tal sentido la mejor separabilidad se da con 7 y 8 clases con valores SC promedio de 0.67 en la BD Rrs_{log}.

Los modelos de aprendizaje automáticos evaluados mostraron una mejor precisión al momento de estimar la concentración de *Chl-a* usando la reflectancia medida por el sensor OLCI (Ocean and Land Colour Instrument) en comparación con el modelo OC_4 usado como referencia. En la fase de validación, el $RMSE$ obtenido usando el método de validación LOOCV mostró una mejora del 6.71%, 3.74%, 4.97%, 0.49% y 2.49% para los modelos PLS, RF, NN, KRR y GPR respectivamente. En la etapa de entrenamiento, donde se valida el modelo con los datos usados para el entrenamiento, mostró una mejora de 24.92%, 32.90%, 38.38%, 91.36% y 89.51% para los modelos PLS, RF, NN, KRR y GPR respectivamente. Los resultados de los experimentos (tabla *8*) mostraron que la transformación de la concentración de *Chl-a* no tiene impacto en los resultados en la fase de entrenamiento y validación. La fase de entrenamiento mejora con el desarrollo de modelos en función de la respuesta óptica del SAMAO, mientras que los mejores resultados en la fase de validación se obtuvieron usando la base de datos completa. Los métodos de aprendizaje automático tipo *kernel* son superiores en la fase de entrenamiento, pero no mejoraron de manera significativa con la fase de validación LOOCV.

Se propone realizar en siguientes estudios una caracterización óptica más completa de SAMAO para crear una clasificación de *tipos de agua* a partir de la cual desarrollar algoritmos adecuados para las especificaciones de cada uno de los tipos ópticos de agua. Los resultados mostrados en la tabla 4 nos dan un primer acercamiento a la separabilidad de los tipos de perfiles ópticos que presenta el agua en SAMAO.

Capítulo III

Caracterización de la dinámica espacio-temporal de florecimientos algales y cambios de color en SAMAO a partir de datos del sensor MODIS entre los años 2003-2020

Resumen del capítulo

Los florecimientos algales en SAMAO se presentan de forma cíclica anual, el crecimiento y posterior decaimiento de estas poblaciones de algas crea cambios de color en el agua. En este capítulo se muestra la evaluación de diferentes algoritmos de clasificación supervisada para identificar los cambios ópticos en el lago usando datos de las bandas del rojo e infrarrojo a 250 m del sensor MODIS en el período de enero 2003 a diciembre 2020. A partir de una revisión de FA registrados en la literatura y análisis estadísticos de gráficos de dispersión, se construyó una base de datos de información espectral y etiquetas del estado de color del lago para evaluar los diferentes algoritmos de clasificación. El mejor clasificador fue Random Forest con una precisión de 87.1%, El análisis temporal y la evaluación espacial de la incidencia de los florecimientos mostraron que mayo, abril y marzo son los meses con mayor presencia de cambios de color en SAMAO relacionados a FA. En el análisis espacial se encontró que la mayor incidencia de florecimientos se da en la región noreste del lago y las mayores cantidades de eventos ocurrieron en los años 2011, 2008 y 2012 respectivamente. Se determina la influencia del fenómeno El Niño-Oscilación del Sur (ENSO) en la incidencia de florecimientos algales en el lago-cráter debido al patrón temporal entre las anomalías en los FA y el índice multivariado de El Niño-Oscilación del Sur, donde el mayor número de eventos de FA se presentaron en las fases frías del ENSO.

Introducción

La presencia y permanencia de FA en los cuerpos de agua continentales genera alteraciones severas en estos ecosistemas, deteriora la calidad del agua generando así problemas de salud pública y llegando a limitar la disponibilidad del recurso (German *et al.,* 2016). Para poder tomar medidas de control y mitigación de los efectos que tienen los FA, es necesario establecer su presencia, temporalidad y los factores que controlan su dinámica (Duan *et al*., 2010). Tradicionalmente, los FA se estudian y caracterizan empleando muestreos rutinarios de la biomasa del fitoplancton a través de la medición de la concentración de *Chl-a* en el agua, junto con la identificación taxonómica de la especie dominante. Sin embargo, los muestreos *in-situ* presentan limitaciones en la accesibilidad, baja cobertura espacial y baja continuidad temporal (Masocha *et al.,* 2018). Es en este aspecto las técnicas de teledetección se presentan como una alternativa complementaria para el monitoreo y estudio de los FA en aguas continentales La firma espectral de *Chl-a* se observa entre las longitudes de onda de 400 a 900 nm, donde hay cuatro bandas (rojo, azul, verde e infrarojo cercano) que se encuentran asociadas con la estimación de la concentración de este pigmento (Rani *et al.,* 2019). Si bien, las primeras estimaciones de la concentración de *Chl-a* satelital fueron desarrolladas en la década de 1970 para ecosistemas oceánicos oligotróficos, el avance de la capacidad computacional y la puesta en órbita de más sensores con mejores resoluciones espaciales y radiométricas han permitido la exploración de nuevos modelos de estimación de *Chl-a* (Xing *et al.,* 2007) en aguas continentales con grados altos de eutrofización. Tradicionalmente se han empleado las bandas azul/verde para aguas abiertas, sin embargo, en aguas más productivas y turbias estas bandas son limitadas para obtener valores de *Chl-a*. Ello se debe a que las concentraciones de partículas

no algales y materia orgánica disuelta coloreada (CDOM del inglés) tienen una fuerte superposición en las características de absorción en el espectro azul. Por lo tanto, en aguas interiores resulta más eficaz aplicar los algoritmos que emplean bandas en el rango de longitud de onda 650-800 nm correspondiente al rojo (RED) e infrarrojo (NIR). Estas bandas son menos sensibles a la absorción por CDOM y dispersión por partículas minerales ya que se desvanecen en gran medida en esas regiones del espectro (Gitelson,1992; Gons, 1999; Schalles, 2006; Dörnhöfer y Oppelt, 2016). En este sentido, se ha desarrollado un gran número de algoritmos para estimar *Chl-a* basándose en la información de las bandas de los espectros NIR y RED, estos algoritmos van desde métodos semianalíticos, relaciones de banda simple, algoritmos multibanda, índices espectrales hasta el aprendizaje de máquina, algunos ejemplos de ello se muestran en la tabla *9*.

El sensor MODIS (Moderate Resolution Imaging Spectroradiometer) lanzado en 1999 y 2002 a bordo de los satélites Terra y Aqua respectivamente, provee de información espectral diaria en diferentes resoluciones espaciales (250 m, 500 m y 1 km). Con sus 36 bandas y cerca de 20 años de serie temporal de datos, MODIS ha probado tener el potencial para monitorear parámetros de calidad de agua en CAC (Wang *et al.,* 2018; Shi *et al.,* 2020; Jia *et al.,* 2019), entre estos la temporalidad, duración y magnitud de FA (Shi *et al.,* 2017, 2019).

La capacidad temporal de MODIS en conjunto con técnicas de aprendizaje automático permiten la creación de nuevos métodos más efectivos en el monitoreo y diagnóstico de FA. Particularmente, los productos MOD09GQ y MYD09GQ de MODIS con resolución espacial de 250 m, han sido ampliamente usados para estudiar FA en CAC (Li *et al.,* 2019; Shi *et al.,* 2015; A. A. Gitelson *et al.,* 2008). Se han empleado estos productos para desarrollar un modelo

de estimación de *Chl-a* aplicado a un embalse de 14 km^2, el modelo obtuvo un R^2 de 0.69 y fue aplicado a una serie de tiempo de datos de reflectividad del período 2001-2014 (cita). Posteriormente Germán *et al.,* (2017) emplearon el algoritmo Harmonic Analysis of Time Series (HANTS) propuesto por Verhoef (1996) para modelar la línea base de patrones estacionales de FA en la serie de tiempo y detectar anomalías en ellos. German *et al.,* (2020) evalúan sus resultados de acuerdo a 4 diferentes enfoques de definición de un FA: a) por límite fijo de Carlson (1977) y b) Tett (1987), c) identificando cambios abruptos en la población de algas midiendo la pendiente de crecimiento rápido y d) según la definición de la Reunión del Consejo Internacional para la Exploración del Mar (ICES, por sus siglas en inglés) 1984 que define los FA como la desviación del comportamiento normal mediante HANTS (Germán *et al.,* 2017). La misma serie de datos 2001-2014 obtuvo diferente número de FA detectados para los cuatro métodos: 816, 303, 270 y 108 respectivamente, dónde los métodos de umbral fijo tendieron a sobreestimar los FA demostrando la importancia de entender las condiciones locales y considerar los patrones estacionales al desarrollar modelos para lagos eutróficos.

En este aspecto, el objetivo de este estudio es caracterizar la dinámica de los FA que se han presentado en SAMAO en el período 2003-2020 usando datos MOD09GQ y MYD09GQ del sensor MODIS obtenidos desde las plataformas espaciales TERRA y AQUA. Para lo cual se caracterizaron las respuestas espectrales de las bandas del rojo e infrarrojo cercano, por medio de clasificación supervisada de eventos registrados en reportes técnicos para evaluar cinco algoritmos de aprendizaje automático, finalmente se procesó la serie completa para identificar los eventos desde la escala estacional a interanual.

Tabla 9. Algoritmos NIR – RED para estimación de Chl-a en aguas continentales

Cita	Algoritmo	Área de estudio	Sensor
Método Semianalítico			
Gons (1999)	Rrs (709/Rrs (665) =>	8 lagos y estuarios	*in-situ*
Gons *et al.* (2008)	aph(665) =>*Chla*	Grandes Lagos E.U.A	*MERIS*
Duan *et al.* (2012)		3 lagos eutróficos	*in-situ*
Relación de banda simple:			
Dekker & Peters (1993)	*Chla* = f (Rrs (709) /Rrs(665))	10 Lagos en Holanda	*in-situ*
Ruddick *et al.* (2001)		Cuerpos de agua en Holanda y Bélgica	*in-situ*
Duan *et al.* (2012)		3 lagos eutróficos en China	*in-situ*
Algoritmo multibanda:			
Dall'Olmo *et al.* (2005)	[Rrs(671) −1-Rrs(710)−1] × Rrs(740) /// [Rrs(662)−1-Rrs(693)−1]× /// [Rrs(740)−1-Rrs(705)−1]	4 lagos y presas en EUA	*in-situ*
Gitelson *et al.* (2008)		Lagos y presas en EUA	*in-situ*
Le *et al.* (2009)		Lago Taihu, China	*in-situ*
Spectral index:			
Mishra & Mishra (2012)	*NDCI, NGRDI, Normalized spectral index*	4 estuarios y bahías, EUA	MERIS
Feng *et al.* (2014)		Lago Poyang, China	MERIS
Shi *et al.* (2015)		Lago Taihu, China	MODIS
Otros métodos: Empirical Orthogonal Function, Machine learning			
Craig *et al.* (2012)	8 bandas en el rango de 320-800 nm	Estación de boya de brújula, Canada	*in-situ*
Qi *et al.* (2014)	4 bandas a 469 nm, 555 nm, 645 nm y 859 nm	Lago Taihu, China	MODIS
Pahlevan *et al.*, (2020)	Neural network: 7 bandas MSI y 12 bandas OLCI entre 400-800 nm	Áreas globales	Sentinel 2 y 3
Prasad *et al.* (2020)	Bandas 440 nm, 560 nm, 655 nm	Ganga River, India	Landsat 8

Metodología

Se realizó un análisis histórico de los eventos de florecimientos algales presentados en SAMAO durante el período 2003-2020. Para este objetivo se procesaron datos de los sensores

AQUA y TERRA a bordo de la misión MODIS empleando técnicas de clasificación supervisada. El proceso se realizó en 4 etapas como muestra la Figura *17*:

i. Exploración: En esta etapa se descargaron datos satelitales diarios del sensor MODIS y se compilaron registros conocidos de eventos de florecimientos algales severos y cambios de color en el lago.
ii. II Caracterización: Se analizó la respuesta espectral de las regiones Red y NIR para 6 clases propuestas: 1. Turquesa, 2. Florecimiento Algal, 3. Homogéneo y 4. Nubes 5. Borde 6. Reflexión especular. (ver tabla *10*).
iii. III. Evaluación: Se evaluaron diferentes técnicas de clasificación supervisada. Se utilizó el módulo de clasificación del programa ARTMO. Se implementó la técnica de validación cruzada leave-one-out, y como métricas para evaluar los resultados de la clasificación se utilizaron matrices de confusión y el índice Kappa.
iv. IV Análisis: Se analizó la distribución espacial y temporal por medio de climatologías entre 2003-2020 para diferentes rangos de coberturas de eventos de FA.

Figura 17. Esquema metodológico para la evaluación de los FA en SAMAO durante 2003 – 2020.

Datos satelitales

El sensor MODIS a bordo de las misiones AQUA (EOS PM-1) y TERRA (EOS AM-1) cuenta con 36 bandas espectrales, la órbita de TERRA recorre un trayecto de norte a sur alrededor de la Tierra, mientras que AQUA pasa de sur a norte. Cada plataforma obtiene una imagen global

de la superficie terrestre cada 1 a 2 días. Las bandas de MODIS van de 0.4 µm a 14.4 µm y al ser diseñadas para cubrir diferentes aspectos de la observación terrestre, su resolución espacial va de 250 m, 500 m y 1000 m. Para este trabajo se descargaron los productos MOD09GQv006/MYD09GQv006 a través de la plataforma: ***Application for Extracting and Exploring Analysis Ready Samples*** (A$\rho\rho$EEARS), estos entregan la reflectancia superficial de las bandas 1 y 2 que abarcan la región del espectro correspondiente a $620nm - 670nm$ y $841nm - 876nm$ respectivamente a una resolución espacial de 250m (Vermote *et al.,* 2015).

Se realizo la re-proyección a coordenadas geográficas Datum WGS84 y EPSG: 4326, y se descargó el producto de banda de calidad correspondiente sugerida por los operadores de MODIS, esta banda fue utilizada para enmascarar píxeles con problemas de lectura en el sensor, presencia de nubes o errores en el proceso de corrección atmosférica. La descarga de los datos se realizó con un polígono formato "shape" que ayuda a descargar únicamente los datos de la región de interés. Los píxeles de la orilla incluyen mediciones de agua y tierra por lo que se creó y empleó una máscara para retirar los píxeles de mezcla de la orilla del lago para evitar el ruido de esta señal, de forma que la matriz de entrenamiento pasó de los 68 píxeles descargados a 43 píxeles de agua, la matriz original descargada en formato shape y la matriz de entrenamiento se muestra en la figura *18*.

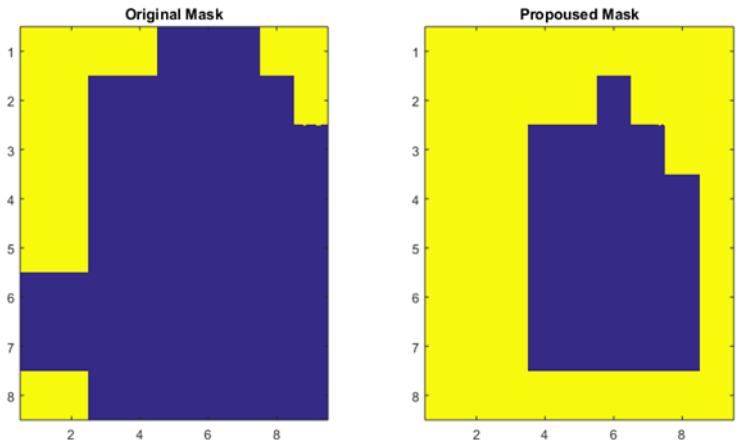


Figura 18. Matriz original y matriz empleada para entrenamiento.

Compilación de eventos FA y determinación de clases

Se desarrolló la base de datos con etiquetas de clase para entrenar el algoritmo de clasificación, en principio esto se realizó a partir de registros de eventos de FA severos y cambios de color totales de SAMAO documentados en estudios previos (Salazar-Alcaraz 2018; Macías 2018), las fechas de estos cambios de color se muestran en la tabla *11*. Posteriormente, con apoyo del visualizador (*https://worldview.earthdata.nasa.gov/*), la segunda parte de las etiquetas se realizó empleando gráficos de dispersión en 2D y 3D, dónde con la visualización del comportamiento de los datos se pudieron ubicar regiones de píxeles pertenecientes a las clases, corroborando las fechas de nuevo con ayuda del visualizador de MODIS. En la figura *19* ejemplifica cómo se visualiza la distribución de los datos, el gráfico contiene los datos diarios NIR-RED de 2003 a 2020 en escala logarítmica, con una tercera dimensión de índice de diferencia normalizada (DNI), la barra de colores indica las fechas de lectura de los datos, cada punto representa un píxel.

Si bien, el fenómeno objeto de este estudio es la dinámica de los crecimientos poblacionales de algas representado directamente por las clases *florecimiento algal* y *turquesa*, se definieron 6 clases para entrenar el algoritmo. Estas clases incluyen los aspectos ópticos del lago que abonan al objeto de estudio: 1. Florecimiento Algal (FA), 2. Turquesa (TQ), 3. Homogéneo (HM); y tres clases más consideradas cómo control para reducir errores en la clasificación: 4. Presencia de nube (NB), 5. Efecto borde de la escena (BR) y 6. Efecto de reflexión especular (SL). La descripción de las seis clases se muestra en la tabla *10*, así mismo, se adjunta un registro fotográfico de las tres clases representativas de los estados ópticos de SAMAO en la figura *20*.

Tabla 10. *Descripción de las 6 clases definidas para los estados de SAMAO*

Clase y acrónimo	Descripción
Turquesa (TQ)	Coloración del agua de un color turquesa 71pecula, generalmente este fenómeno ocurre después de un FA severo (figura *20a)*
Florecimiento Algal (FA)	Zonas donde 71pecula presenta severos crecimientos en las 71pecular71es de microalgas en su superficie. Período la coloración verde 71pecula y la profundidad óptica del lago es muy baja (figura *20b)*
Homogéneo (HM)	Estado habitual del agua en 71pecula vista desde 71pecula remotos. El tono 'oscuro' homogéneo se mantiene durante la mayor parte del año (figura *20c)*
Nubes (NB)	Esta etiqueta 71pecula eventos que corresponden a una alta densidad de nubes.
Borde (BR)	Píxeles con distorsión 71pecular debido al borde del Swath (ancho del área que 'barre' el sensor, 2300 km de ancho)
Reflexión 71pecular (SL)	Píxeles con efectos de reflexión 71pecular por la geometría de observación del sensor y el ángulo del Sol.

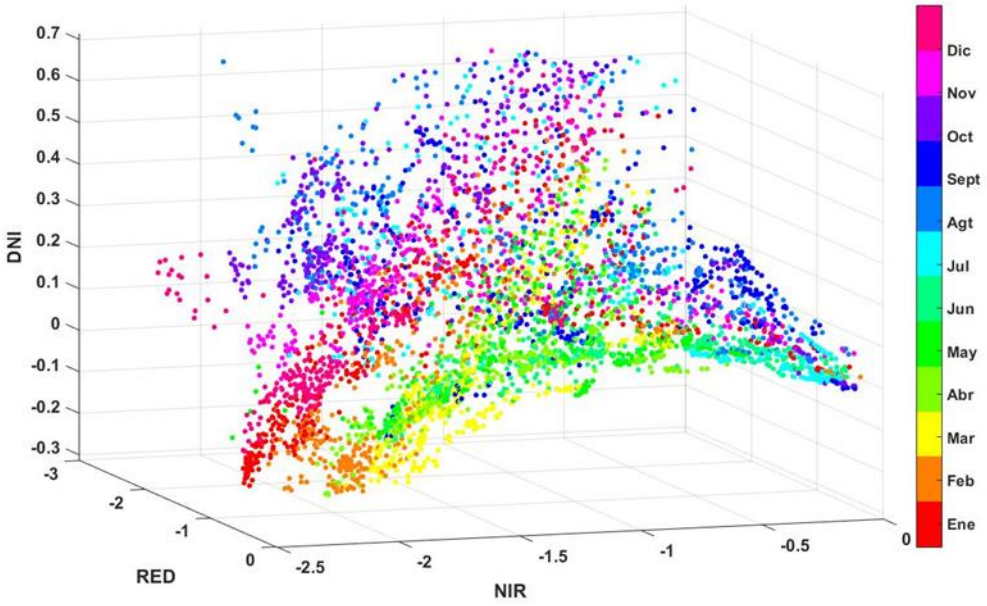


Figura 19. Gráfico de dispersión en 3D con datos diarios NIR, RED y DNI de la serie 2003-2020.

Tabla 11. Fechas documentadas de cambios de color totales y florecimientos algales severos en SAMAO.

TURQUESA		
Día	Mes	Año
7	5	2015
3	4	2018
30	3	2018
2	5	2020
21-23	11	2020
FLORECIMIENTO ALGAL		
Día	Mes	Año
9	3	2018
14	3	2018
24	3	2018
12	3	2018

(a)

(b)

(c)

Figura 20. Cambios de color en Santa María del Oro. A) Turquesa, (julio 2018); b) Presencia de florecimiento algal (marzo 2018) Salazar-Alcaraz (2018); c) Homogéneo, febrero 2020.

Para complementar la base de datos de los registros de eventos FA se desarrolló un algoritmo para búsqueda de eventos a lo largo de la serie temporal analizada. Esta aplicación presenta diversas funciones como:

- Visualización multidimensional y multiescala en 2D y 3D y en escala normal o logarítmica de los datos, como lo muestra el ejemplo en la figura *19*.
- Visualización de eventos en la plataforma de la NASA Worldview (*https://worldview.earthdata.nasa.gov/*)
- Selección de píxeles y caracterización de las bandas del rojo y NIR.
- Generación de bases de datos en formato de texto.

Caracterización de las clases

La caracterización de las clases en la base de datos se realizó, en principio, usando diagramas de cajas (Williamson *et al.,* 1989) éstos nos dan representación gráfica de la distribución y simetría de cada atributo clasificador usado. En la figura *21* muestra un ejemplo de la visualización de los datos en $\log_{10}$ para las clases: TQ, FA y HM. Para entrenar un algoritmo de clasificación es necesario atribuir a cada dato una serie de características que permita distinguirlos entre clases mejor definidas. Dos parámetros en cada píxel son los valores Rrs en las bandas Red y NIR de MODIS, además, se añadieron 3 relaciones espectrales como variables clasificadoras (tabla *12*). Así, tenemos una base de datos con 5 parámetros asociados a cada píxel para aumentar la precisión en el entrenamiento del algoritmo, estos se enlistan a continuación:

i. Banda 1 $(620nm - 670nm)$ [red]

ii. Banda 2 $(841nm - 876nm)$ [NIR]

iii. Simple Ratio 1 $\frac{NIR}{red}$ [SR1]

iv. Simple Ratio 2 $\frac{red}{NIR}$ [SR2]

v. Diferencia Normalizada $\frac{NIR-red}{NIR+red}$ [DNI]

La Tabla *12* muestra de manera más detallada las relaciones espectrales empleadas como atributos para la clasificación, se muestra en la primera columna el nombre genérico del índice seguido de su formulación matemática y referencia bibliográfica.

Tabla 12. Relaciones empíricas de 2 bandas basadas en el NIR-red para estimar concentración de clorofila

Índice espectral	Ecuación	Fuente
Proporción simple	$\frac{NIR}{red}$ $\frac{red}{NIR}$	Moses *et al.*, 2009; Huang *et al.*, 2014; Lyu *et al.*, 2015
Diferencia normalizada	$\frac{NIR - red}{NIR + red}$	Mishra & Mishra, 2012

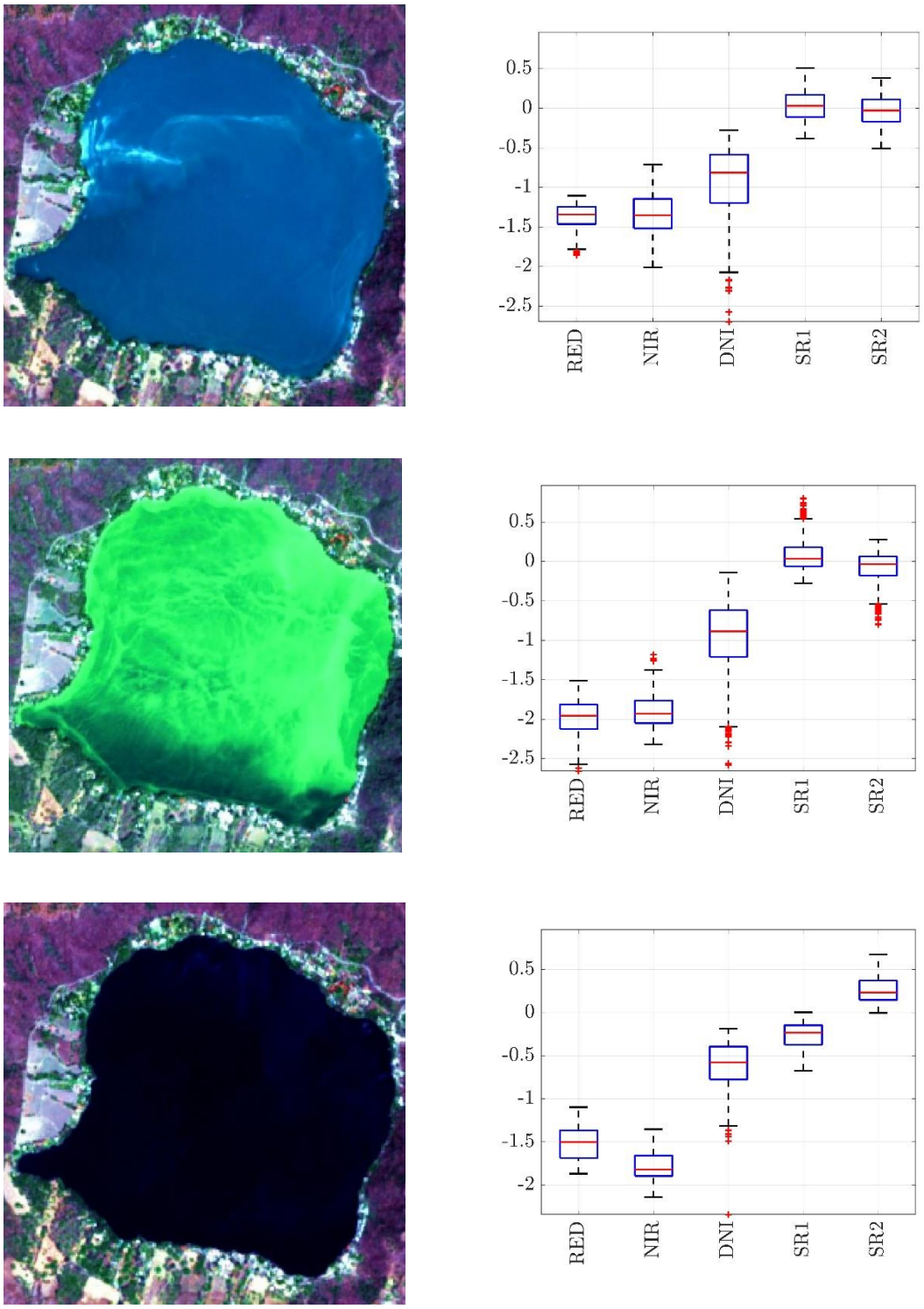


Figura 21. Valores de las variables clasificadoras en log_{10} para las clases de cambio de color. Ejemplo con recortes de Sentinel 2 MSI.

Algoritmos clasificadores

Se evaluaron los algoritmos de clasificación supervisada: 1. Análisis Discriminante, 2. K-vecinos cercanos, 3. Naive Bayes y 4. Random Forest, y 5. Redes Neuronales. Estos algoritmos han sido implementados en: the Simple Classification Toolbox (Muñoz & Camps, 2013 (accessed October 21, 2020)) desarrollada por el Grupo de Observación de la Tierra (LEO) de la Universidad de Valencia (España) y que ha sido implementada en el módulo de clasificación de programa ARTMO toolbox (Verrelst & Rivera, 2019 (accessed May 11, 2022)). Para la validación se empleó la estrategia de $Kfold - n$ con un total de 10 subconjuntos.

A continuación se da una breve descripción de los algoritmos evaluados en este trabajo:

Análisis Discriminante

Es una generalización del discriminante lineal de Fisher (Fisher 1936) usado en clasificación multivariada y perimétrico que asume que los datos de las diferentes clases tienen una distribución normal. Se realizó un análisis discriminante lineal el cual genera un modelo que tiene la misma matriz de covarianza de cada clase variando solo la medias. el objetivo es minimizar la función:

$$\hat{y} = \ arg\ min \sum_{k=1}^{K} \hat{P}(k|x)\ C(y|k)$$

donde $\hat{y}$ es la clase que se estima al punto x, K es el numero de clases, $\hat{P}(k|x)$ es la probabilidad a posteriori de que el punto x pertenezca a la clase k y $C(y|k)$ es el costo de clasificar el punto x como y cuando su verdadera clase es k. (Tharwat *et al.,* 2017) entrega una descripción detallas del modelo de análisis discriminante.

Vecinos cercanos

Este algoritmo no perimétrico asigna la clase $\hat{y}$ a un punto x por el máximo números de k a partir de una función de costo C la cual calcula métricas de distancia entre el nuevo punto a clasificar y los datos usados en el entrenamiento del modelo (Fix & Hodges, 1989). La función de costo seleccionada fue Mahalanobis el cual utiliza la matriz de covarianza de los datos de entrenamiento.

Dado un conjunto de entrenamiento X de n puntos y la función de costo C, knn busca los k vecinos cercanos al punto y_t asignando la clase mayoritaria de los k vecinos.

$$d_{st}^2 = (x_s - y_t)C^{-1}(x_s - y_t)'$$

donde x_s es el vector s del conjunto de datos X, y_t vector de características que se desean clasificar y C es la matriz de covarianza.

k-nn es tambien denominado algoritmo vago porque todo el cálculo lo realiza en la etapa de clasificación.

Naive Bayes

Este algoritmo se fundamenta en el teorema de Bayes (Flach & Lachiche, 2004), es un clasificador probabilístico con algunas simplificaciones en sus supuestos teóricos sobre la independencia de las variables predictoras dadas las clases, por lo cual tiene el sobrenombre de *ingenuo (naives)*

Este algoritmo calcula la probabilidad de pertenecía de un nuevo punto a cada clase del entrenamiento y le asigna la de mayor probabilidad según la siguiente formula:

$$\hat{P}\big(Y = k|X_1, \ldots, X_p\big) = \frac{\pi(Y = k)\prod_{j=1}^{P} P\left(X_j|Y = k\right)}{\sum_{k=1}^{k} \pi\left(Y = k\right)\prod_{j=1}^{P} P\left(X_j|Y = k\right)}$$

donde Yes la variable aleatorio correspondiente a la a clase, $X_1, \ldots, X_p$ son predictores aletorios de un conjunto de datos. y $\pi(Y = k)$ es la prioridad a priori que la clase pertenezca a k

Random Forest

Random Forest o Arboles aleatorio es un ensamble de varios árboles de decisión que han sido entrenados de manera paralela (Breiman 2001). El valor de la clasificación de una nueva muestra corresponde a la clase mayoritaria que han entregado el ensamble de los árboles. Las reglas de decisión se basaron en el índice de impureza Gini (Raileanu & Stoffel, 2004) que mide el grado de "impureza" de un nodo: $Gini(t)$ iguales a cero indican que los datos que pertenecen a una sola categoría, mientras que índices mayores que cero y con valores hasta de uno indican nodos donde los datos pertenecen a más de una categoría. El índice esta definidos así:

$$Gini(t) = 1 - \sum_{i=1}^{j} P\left(i|t\right)^2$$

donde j representa el numero de clases en la etiqueta y P representa la relación de clases en el i_{th} nodo.

Redes Neuronales

Se utilizó una red neuronal prealimentada (feed-forward en inglés) donde las conexiones entre las unidades no forman un ciclo, la información se mueve en una únicamente hacia adelante. Para su diseño y entrenamiento se utilizó la función *patternnet* de la Deep Learning Toolbox de

Matlab (The MathWorks 2010). Su topología es de cuatro capas: 1. Nodos de entrada, 2. Capa oculta, 3. Capa de salida y 4. Salida. La figura *22* muestra la topología clásica de esta red neuronal. Las características de entrenamiento de esta red neuronal son:

- Algoritmo de entrenamiento: Scaled Conjugate Gradient (SCG) con tasa de convergencia superlineal (Roodschild *et al.,* 2019).
- Función de minimización: Cross-Entropy, la cual se utiliza para para ajustar los pesos de la red neuronal en la capa *output* (Q. Wang *et al.,* 2022).

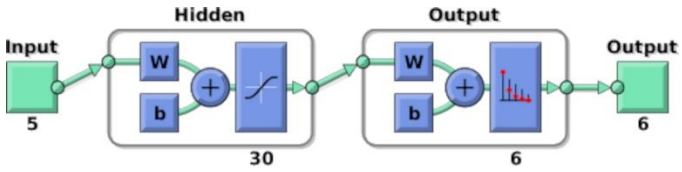


Figura 22. Topología tipo de una red neuronal para el reconocimiento de patrones usada por la función patternnet.

Evaluación de la precisión

Los algoritmos de clasificación supervisada se evaluaron usando las métricas de Overall Accuracy (*a*), para el análisis a nivel de clases se calcularon las métricas de Precision (*b*) como estimador de relevancia y Recall (*c*) como métrica de sensibilidad de los algoritmos de clasificación.

$$\text{a)}\quad \text{Overall accuracy} = \frac{T_p+T_n}{T_p+T_n+F_p+F_n}$$

$$\text{b)}\quad \text{Precision} = \frac{T_p}{T_p+F_p}$$

$$\text{c)}\quad \text{Recall} = \frac{T_p}{T_p+F_n}$$

dónde T_p es el número de positivos verdaderos,T_n es el número de negativos verdaderos F_p es el número de falsos positivos, y F_n el número de los falsos negativos.

Análisis espacial y temporal de cambios de color

Se analizó a escala espacial y temporal la frecuencia de ocurrencia de los siguientes eventos:

- Florecimientos Algales
- Cambios a Turquesa
- Estado Típico

Para la escala espacial se realizó el análisis a nivel del píxel por medio del mapeo de las frecuencias de ocurrencia de los eventos y para la escala temporal se analizaron los promedios mensuales y medias climatológicas.

Se definieron 3 escenarios para el análisis en función del porcentaje de cobertura del área de estudio:

i. Baja cobertura [BC]: Se analizan si por lo menos existe un 30% de la cobertura del evento.
ii. Media cobertura [MC]: Se analizan si por lo menos existe un 50% de la cobertura del evento.
iii. Alta cobertura [AC]: Se analizan si por lo menos existe un 70% de la cobertura del evento.

Resultados

Base de datos

Para entrenar los algoritmos de clasificación se construyó una base de datos con 3529 observaciones (BDclass), donde cada observación corresponde a un píxel etiquetado de acuerdo a los criterios de determinación de clases. BDclass se encuentra equilibrada, cada clase cuenta

con 588 observaciones, los boxplots en la figura 23 muestran el comportamiento de los valores obtenidos en los índices clasificadores (eje *Y*) correspondientes a cada clase (eje *X*). Estos valores son usados como entrada al modelo de clasificación para caracterizar y definir cada una de las seis clases en la fase de entrenamiento.

Las reflectividades en las bandas RED y NIR presentan comportamiento semejante entre sí, con una dispersión alta en las clases de control y baja en las clases de color del agua. Estos resultados son los esperados debido al alto albedo connatural de las nubes y los efectos que engloban las clases NB, SL y BR, mientras que las clases FA, TQ y HM tienen baja dispersión en sus valores debido al carácter sumamente absorbente del agua que contribuye con solo el 20% de la radiación total que mide el sensor (Moses *et al*., 2017).

Clasificación

Se llevó a cabo la evaluación del rendimiento de los algoritmos de clasificación. Los algoritmos de AD, KNN y NB se evaluaron a partir de la función de optimización (hyperparameters). Para la evaluación de AD se optimizaron los parámetros: Delta y Gamma, y se obtuvo un valor *OA* del 65.5%. Para KNN se ajustaron los parámetros de número de vecinos y métricas de distancia obteniendo un 81.3% de *OA*. Para NB se ajustaron los parámetros de tipo de distribución (normal, kernel) y los pesos datos a las muestras, éste obtuvo un 76% de *OA*. Para NN se evaluaron diferente número de nodos en la capa oculta: 2, 4, 6, 10, 15, 20, 25 y 30, donde los mejores resultados fueron de 70.8% *OA* obtenidos con 30 nodos. De modo que la topología de NN consistió en 5 entradas, 30 capas ocultas, 6 capas de salida y 6 salidas. Para Random Forest se evaluaron diferentes números de árboles entre 5 y 400, el mejor desempeño se obtuvo con un bosque de 20 árboles con valores de *OA* de **83.3%**.

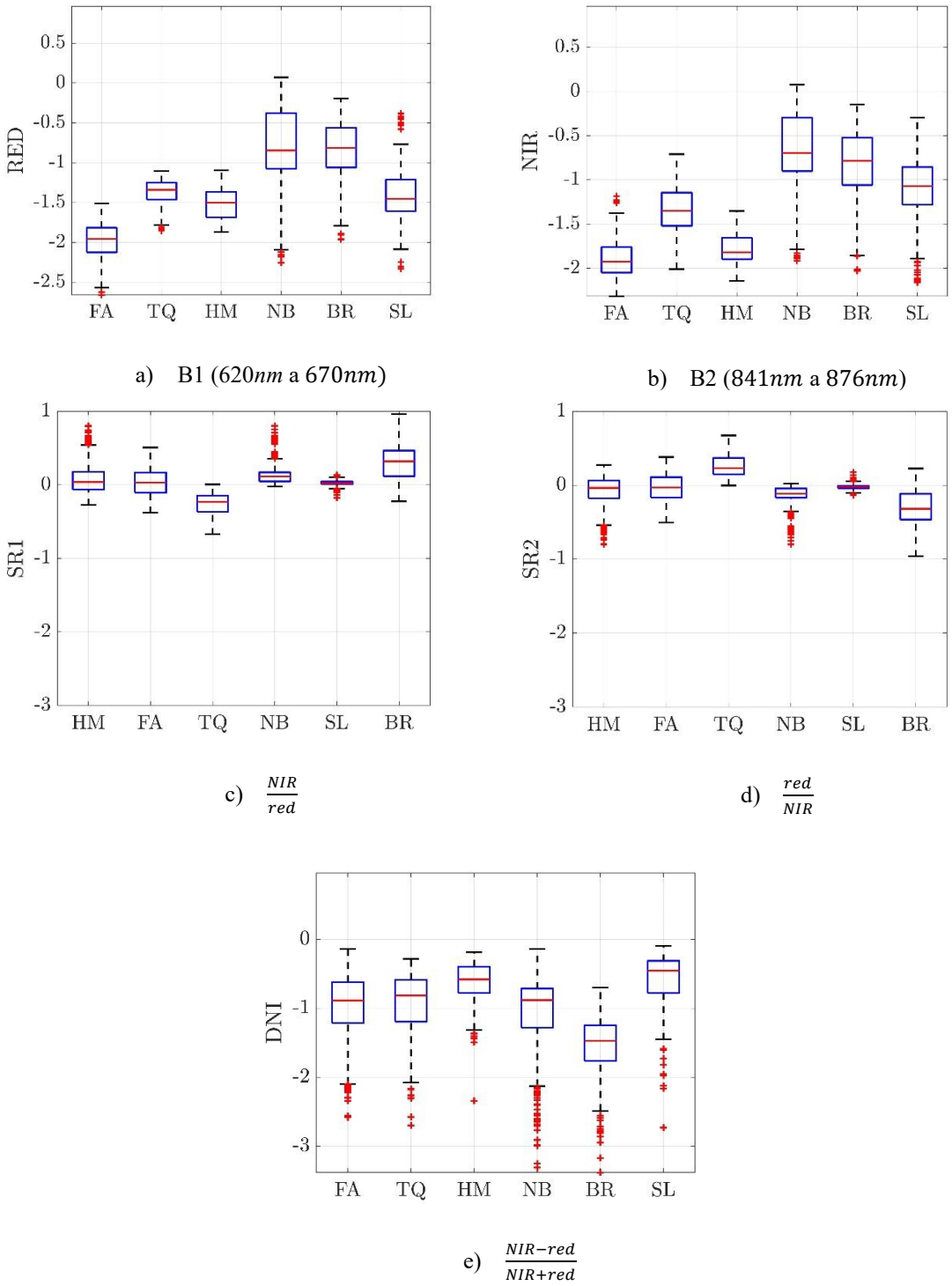


Figura 23. Diagrama de bigotes de la distribución de las clases en función de cada variable clasificadora.

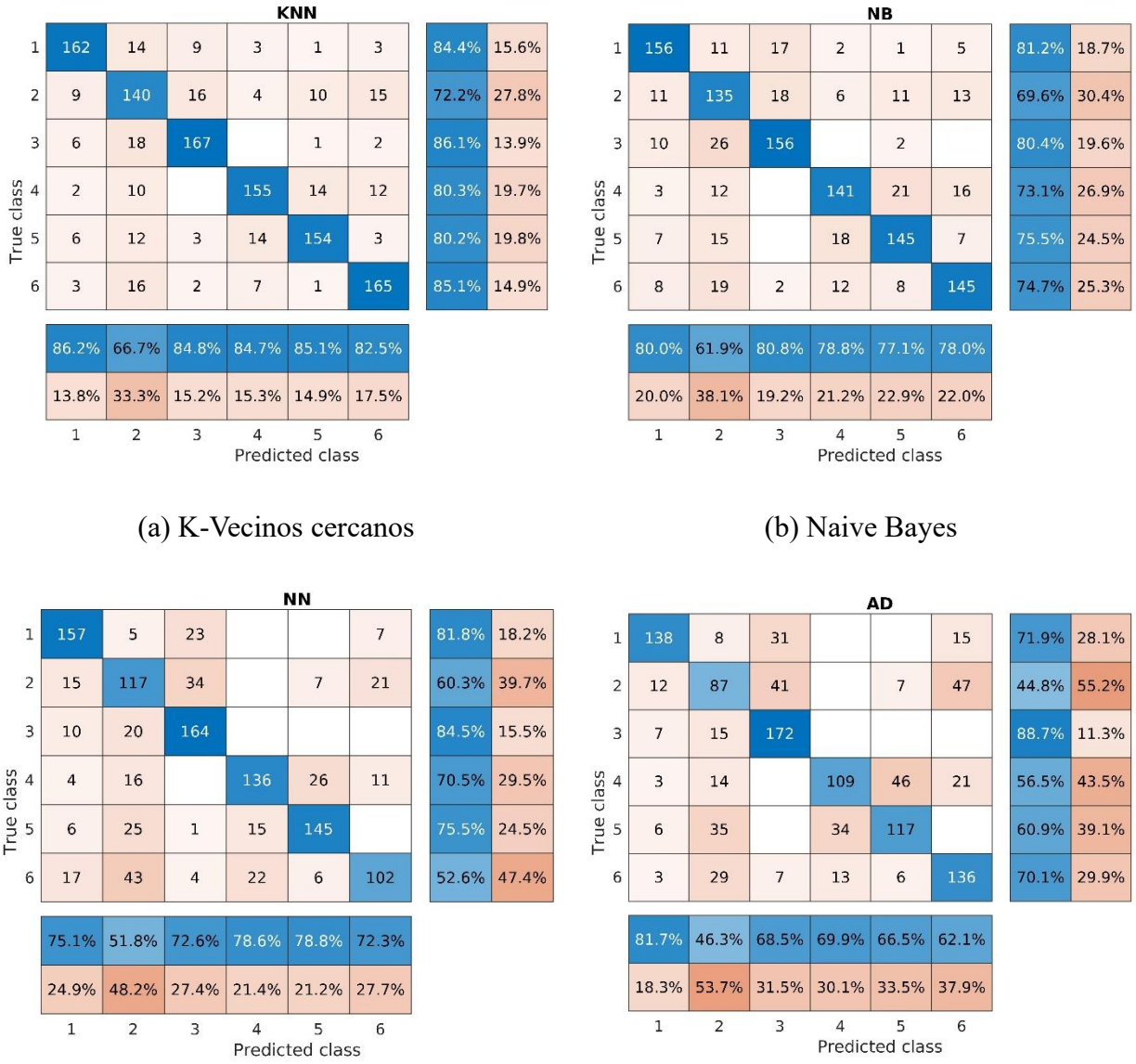
En la tabla 13 se muestran los resultados de las métricas *Overall Accuracy*, *Precision* y *Recall* para el conjunto de algoritmos evaluados, y en la figura 24 se detalla mediante matrices de confusión el desempeño de éstos. Imprevistamente, los mayores errores en la clasificación ocurrieron para la clase TQ en la evaluación de los 5 algoritmos, con mayor número de falsos positivos en la clase HM con KNN y NB, y en la clase BR con RF y AD.

Tabla 13. Métricas del desempeño de los algoritmos clasificadores evaluados.

Algoritmo	**Overall Accuracy**	**Precision**	**Recall**
Análisis Discriminante	65.49	65.8	65.5
K-vecinos cercanos	81.36	81.6	81.4
Naive Bayes	76.01	76.6	76
Random Forest	**83.35**	83.6	83.3
Redes Neuronales	70.84	71.5	70.9

El algoritmo KNN obtuvo resultados *OA* de apenas 1.99% menores a RF, por lo que se considera tener potencial para clasificar los estados ópticos de SAMAO, mientras que AD tuvo errores mayores al resto de algoritmos. RF ha probado ser eficaz como clasificador en estudios de FA previos (Pal, 2005; Shaik y Srinivasan, 2019). Ananias *et al.,* (2022) integraron RF a una cadena automatizada de detección de FA mediante datos MODIS, donde en la fase de evaluación el algoritmo obtuvo un acierto del 95%. La figura 25a muestra las simulaciones realizadas con diferentes bosques conformados por entre 5 y 400 árboles, de acuerdo a Breiman (2001) el desempeño de RF mejora con el incremento de árboles en el bosque. Sin embargo, en nuestros resultados indican que RF con un tamaño de bosque de 20 árboles presentó el mayor valor de *OA*. Los resultados de precisión en la matriz de confusión de RF (figura 25b) para las clases FA,

TQ y HM fueron de 82.8%, 71.9% y 83.9% respectivamente. Las clases de control NB, BR y SL por su parte obtuvieron 87.4%, 88.1% y 87.3% de precisión.



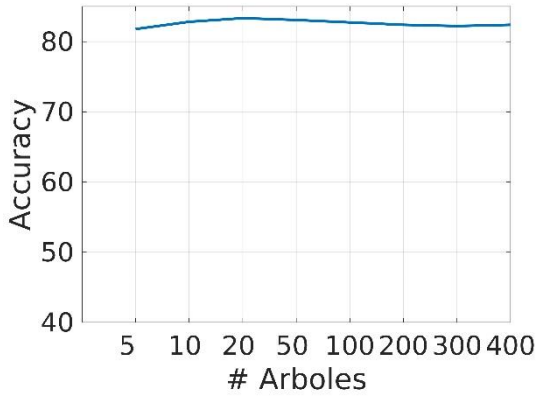
(a) K-Vecinos cercanos

(b) Naive Bayes

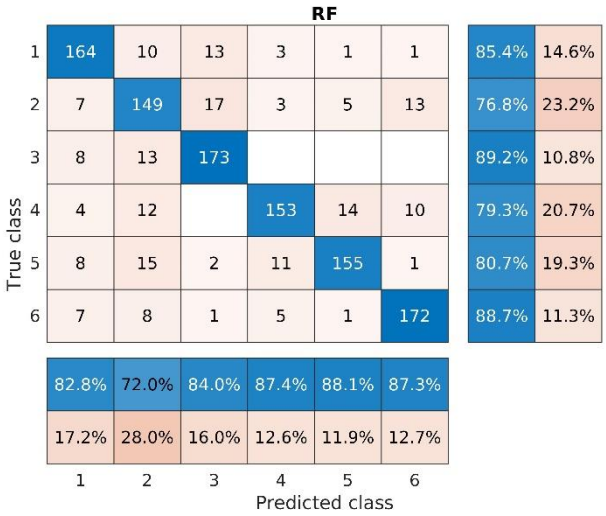
(c) Redes Neuronales

(d) Análisis discriminante

Figura 24. Matrices de confusión en la evaluación de los algoritmos KNN, NB, NN y AD.



a) Relación de la precisión en relación al número de árboles que conforman el bosque.



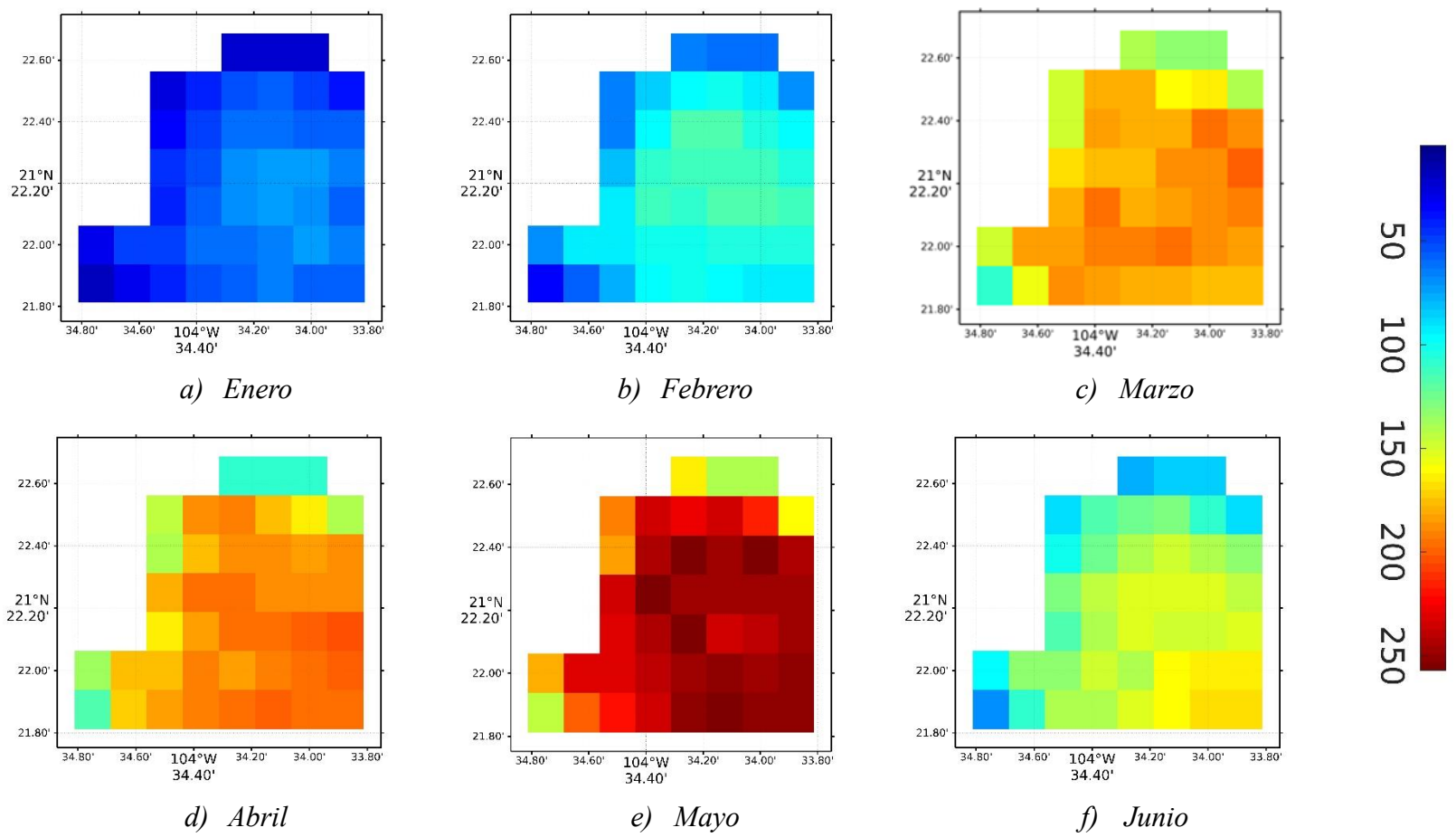
b) Matriz de confusion.

Figura 25. Desempeño de Random Forest.

Caracterización espacial y temporal: escala mensual

Se realizó la caracterización espacial y temporal 2003-2020, presentamos los resultados obtenidos para la clase FA. En orden de mostrar la variabilidad espacial de los florecimientos los resultados se representan a través de matrices de SAMAO (figura 26), éstas corresponden a

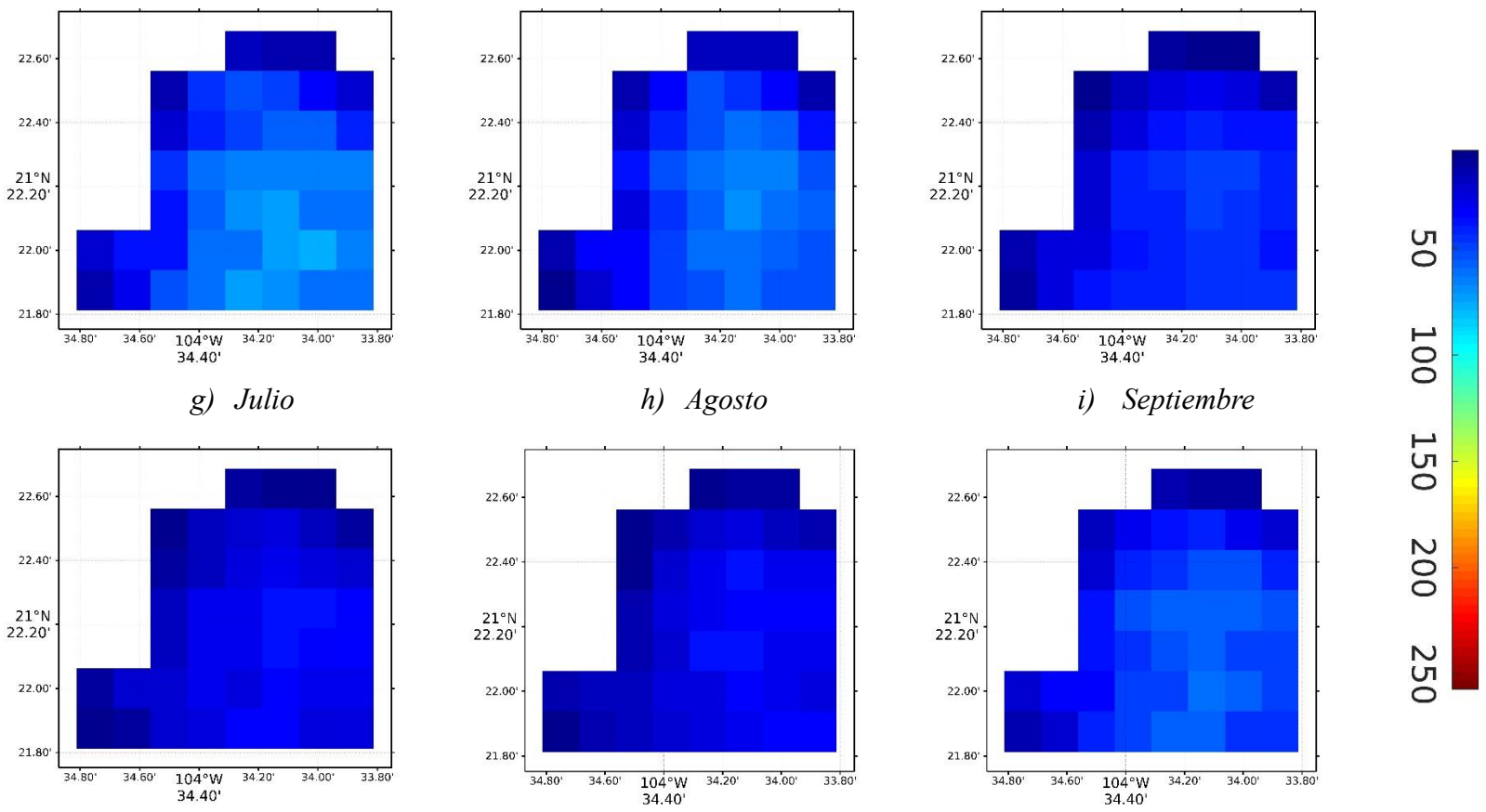
la máscara de píxeles empleada para la extracción de datos, la barra de valores que va de 0 a 250 indica el número acumulado de ocasiones en que *cada píxel* fue clasificado como FA en la serie tiempo. Los eventos de FA varían mensualmente iniciando el incremento de febrero (100 eventos) hasta el máximo en Mayo (250 eventos), descendiendo hasta julio (<50 eventos). De julio a diciembre las frecuencias de FA se mantienen bajas. Este comportamiento temporal corrobora la periodicidad cíclica de los FA en SAMAO, asociados a la estación de primavera. Espacialmente, los resultados indican que los píxeles con mayor incidencia de FA se ubican en la región sureste del lago, lo cual asociamos como una respuesta al régimen diurno del viento, el cual es uno de los principales componentes en la distribución espacial superficial de las poblaciones de algas en los CAC (Hsiao, 1988). En este sentido, las características orográficas de SAMAO definen un patrón en las fluctuaciones del viento donde, la brisa diurna se mueve al oeste por la noche y al este por la tarde con una rotación en sentido antihorario. Alrededor del mediodía, en las horas aproximadas al paso del sensor, Serrano *et al.,* (2002) encontraron que se forma un giro anticiclónico en la parte sur del lago lo que puede explicar por qué estos píxeles han sido los de mayor incidencia de FA detectados.



a) Enero *b) Febrero* *c) Marzo*

d) Abril *e) Mayo* *f) Junio*

Figura 26. Incidencia espacial de FA en SAMAO por períodos mensuales para la serie de tiempo 2003-2020. La barra de valores indica el número acumulado de ocasiones en que cada píxel fue clasificado como FA en la serie de tiempo.



g) Julio *h) Agosto* *i) Septiembre*

j) Octubre *k) Noviembre* *l) Diciembre*

Figura 26. Incidencia espacial de FA en SAMAO por períodos mensuales para la serie de tiempo 2003-2020. La barra de valores indica el número acumulado de ocasiones en que cada píxel fue clasificado como FA en la serie de tiempo (Cont.).

Se determinó como un *evento de FA* a aquellas fechas donde el área de píxeles clasificada fuese superior al 30% de la cobertura del lago. La figura 27 muestra el acumulado de eventos de FA obtenidos entre los años 2003-2020 a escala mensual, se observa un patrón estacional dónde mayo, abril y marzo son, en ese orden, los meses con la mayor acumulación de eventos, lo que coincide con los resultados encontrados en 2015 por Cortés-Macías, (2018) y Salazar-Alcaraz *et al.,* (2021), donde el florecimiento tuvo lugar de enero a abril. Los meses con menor incidencia de FA encontrados ocurren en enero y de julio a diciembre, cuando como indican Sosa-Nájera *et al.,* (2010) el lago presenta alta estratificación. Germán *et al.,* (2016) encontraron un comportamiento periódico en los FA con picos en verano y valles en invierno empleando una BD MODIS del período 2001-2014, Shi *et al.,* (2019) evaluaron un modelo NIR-RED con datos de entre 2013-2017 de MODIS y reportaron también ciclos estacionales con picos anuales de FA en julio-agosto.

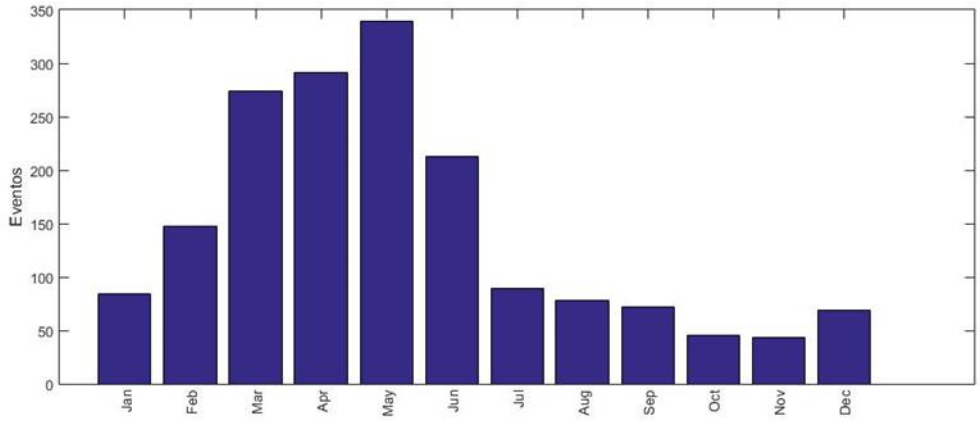
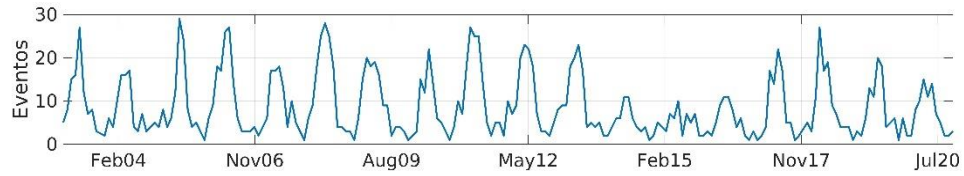
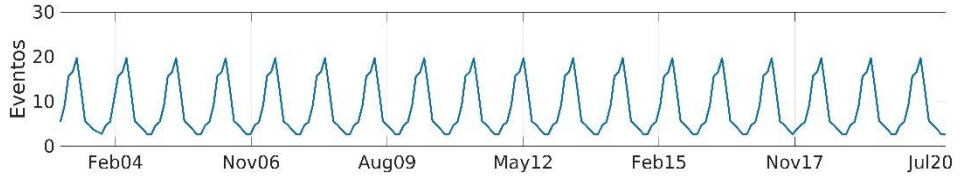


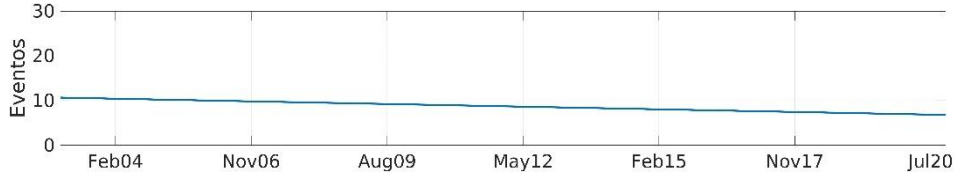
Figura 27. Ocurrencia de eventos FA a escala mensual.



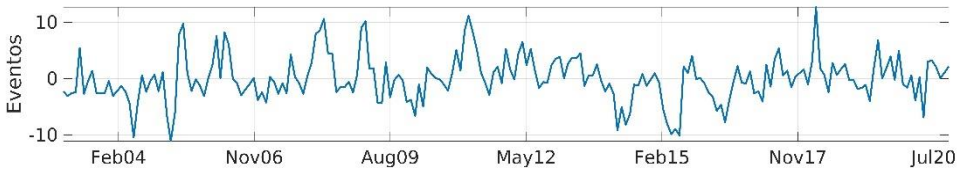
(a) Serie de FA 2003-2020



(b) Ciclo estacional



(c) Tendencia



(d) Anomalías

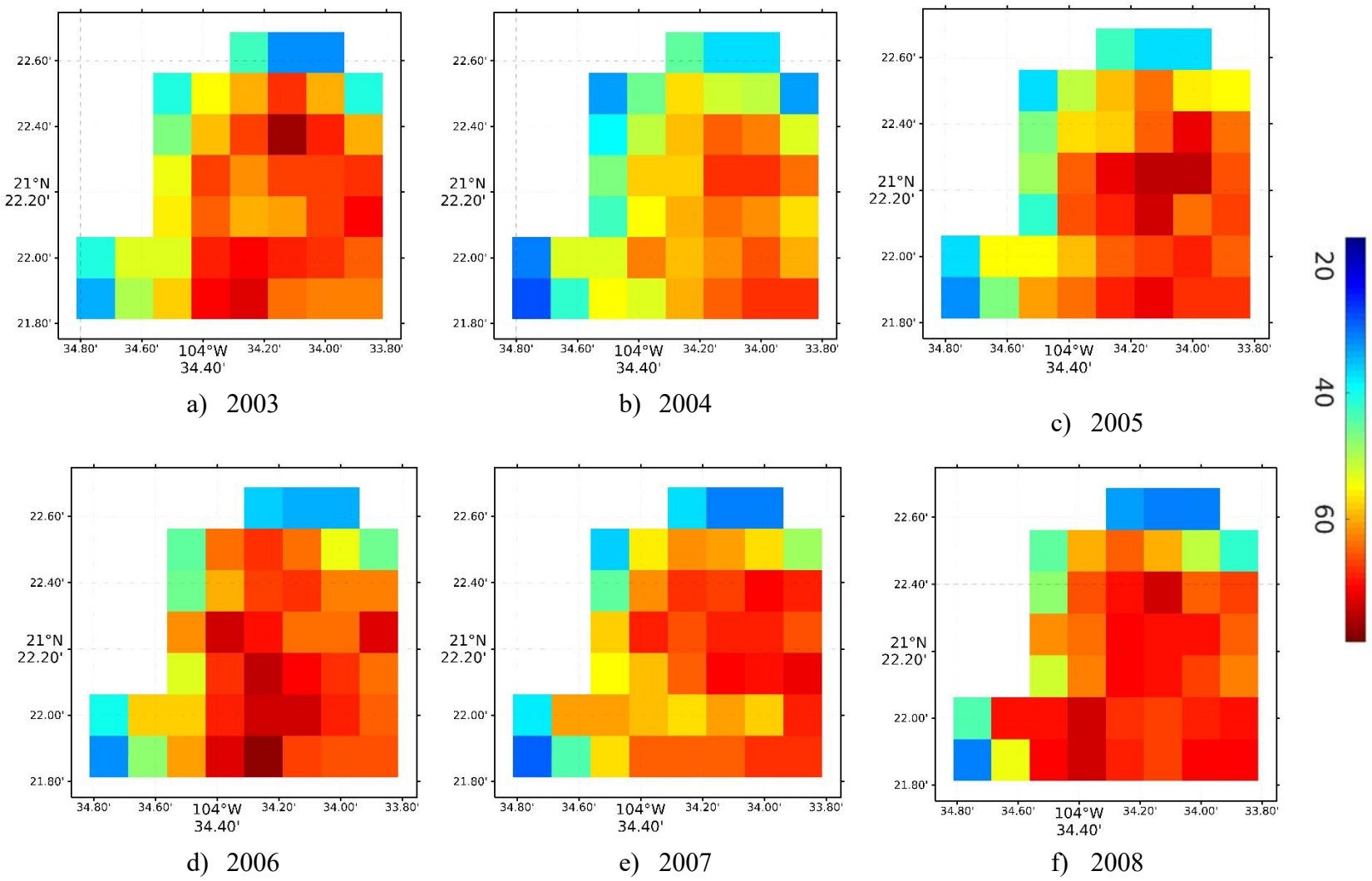
Figura 28. Descomposición de la serie de tiempo de FA obtenidos a partir del algoritmo de clasificación RF empleando datos MODIS, lago SAMAO, 2003-2020.

La Figura 28 muestra el análisis de la descomposición de la serie temporal de los eventos FA durante el período 2003-2020 agrupados mensualmente. Dónde la figura 28a muestra la distribución observada del número de eventos mensuales que se presentaron en el período de estudio; el ciclo estacional de los FA se extrajo de la serie temporal (figura 28b) donde el patrón temporal presenta un marcado ciclo anual, con un máximo en mayo y un mínimo en octubre indicando que es la escala temporal dominante en el comportamiento de los FA;

Complementariamente, la variabilidad interanual indica entre los valores observados (figura 28a) y el ciclo anual (figura 28b), indican la influencia de fenómenos de diferentes escalas actuando sobre SAMAO. la figura 10c indica la tendencia de los eventos la cual se ha ajustado a una recta del tipo y=mx+b, donde la pendiente (m) tienen un valor de -0.00052 y el corte con el eje y de 392.53. El valor de la pendiente nos indica que se observa una tendencia negativa muy baja en la serie; la figura 28d nos muestra las anomalías: una resta del ciclo estacional a la serie completa, dichas anomalías abarcan un rango entre -10 a 11.

Caracterización espacial y temporal: escala anual

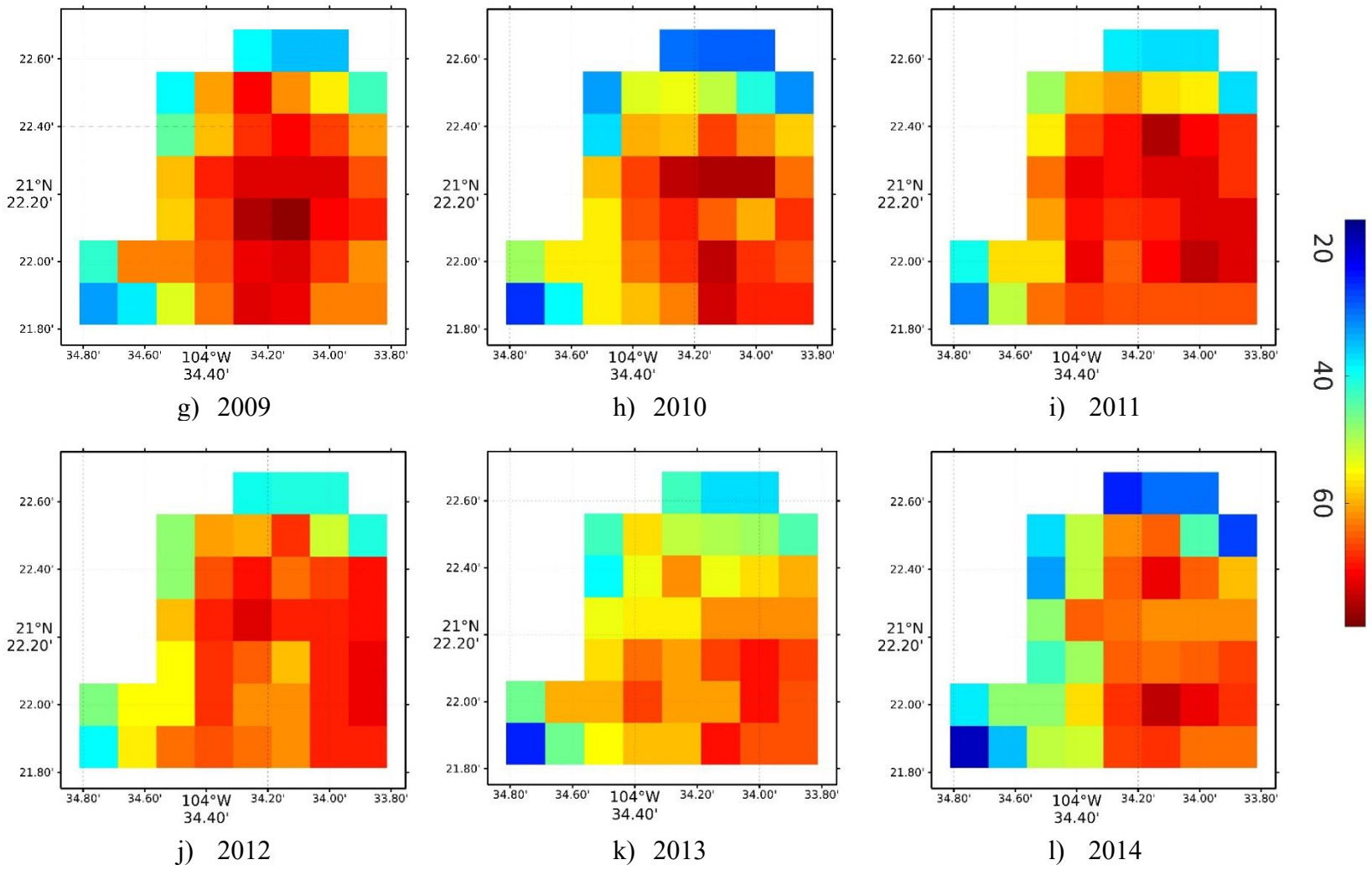
La caracterización espacial a escala anual de los florecimientos algales en SAMAO se muestra empleando las matrices de la figura 29. Las matrices representan la máscara de píxeles empleada en este trabajo. La barra de valores indica el número de días (eventos) en que cada pixel de la matriz de datos (área de aprox. 250 m2) tuvo valores de reflectancia etiquetados como ‘florecimiento’. Observamos que los años 2011 y 2008 tienen la mayor cantidad de píxeles con incidencia de eventos FA arriba de 70. Mientras que 2004 y 2015 tienen en la mayoría de sus píxeles una incidencia de eventos por debajo de 60.



a) 2003 b) 2004 c) 2005

d) 2006 e) 2007 f) 2008

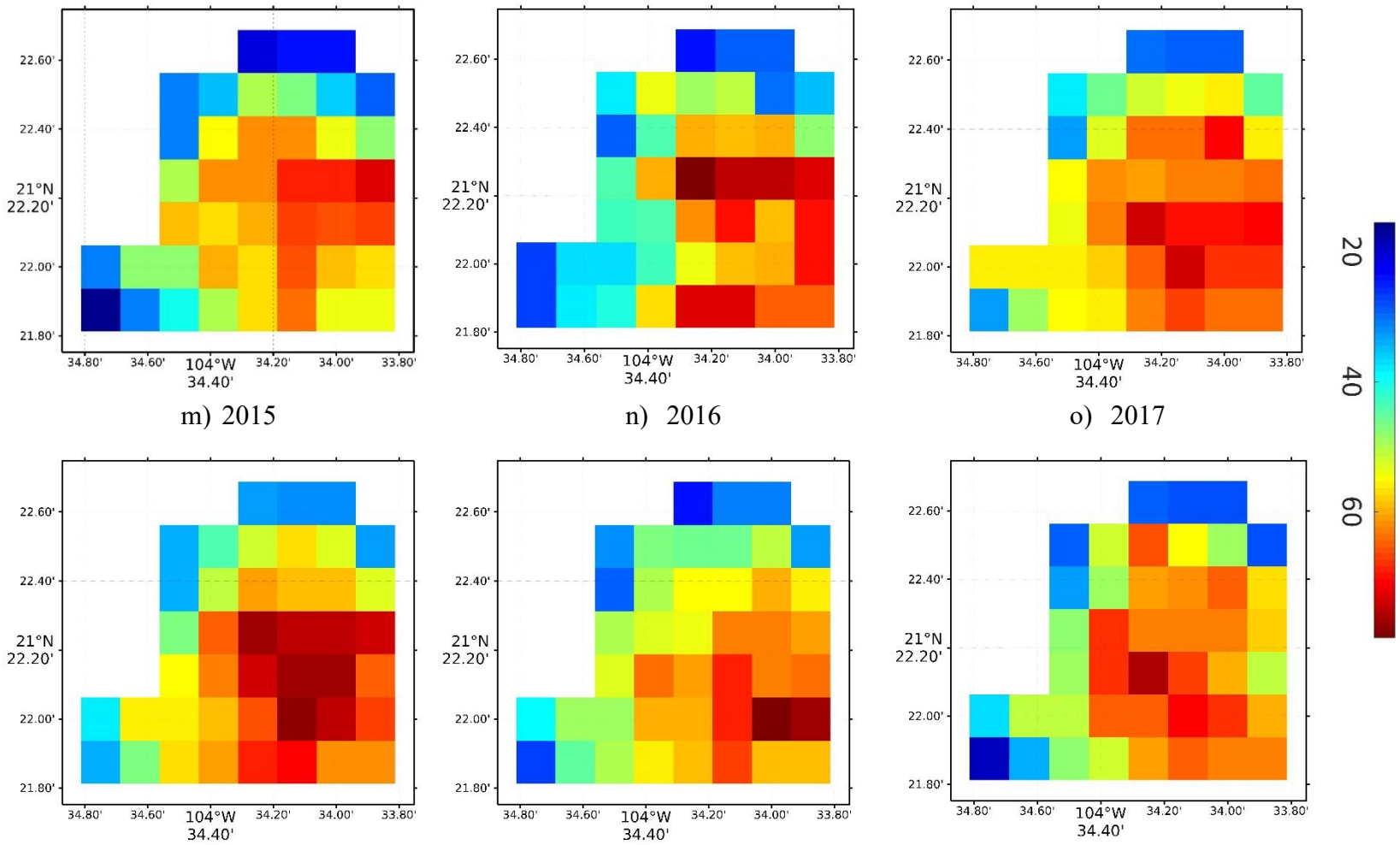
Figura 29. Incidencia espacial de florecimientos algales en el lago por períodos anuales para la serie de tiempo 2003-2020.



g) 2009 h) 2010 i) 2011

j) 2012 k) 2013 l) 2014

Figura 29 Incidencia espacial de florecimientos algales en el lago por períodos anuales para la serie de tiempo 2003-2020 (Cont.).



m) 2015 n) 2016 o) 2017

p) 2018 q) 2019 r) 2020

Figura 29 Incidencia espacial de florecimientos algales en el lago por períodos anuales para la serie de tiempo 2003-2020 (Cont.).

Como se observa en la figura *30*, los años 2011, 2008 y 2012 son, en ese orden, los que mayor número de eventos de FA presentaron, estos años coinciden con períodos niña del ENSO. Mientras que los años 2015, 2016 y 2014, que corresponden a períodos niño, tienen los menores registros de eventos clasificados como FA. En la figura *31* se observa que existe una tendencia inversa de las anomalías de FA obtenidas en este trabajo con el índice multivariado de El Niño/Oscilación del Sur (MEI) ('https://psl.noaa.gov/enso/mei/data/meiv2.data'). De modo que anomalías positivas de FA obtenidas en este trabajo (período 2003-2020) ocurren durante períodos fríos del ENSO y anomalías negativas de FA en SAMAO ocurren en períodos cálidos de ENSO. En la figura 31 se representan los periodos cálidos (barras en color rojo) o fríos (barras en color azul) que sobrepasan el umbral de +0.5°C o -0.5°C e indican la ocurrencia de El Niño o La Niña respectivamente, las anomalías de eventos FA se representan con una línea sólida.

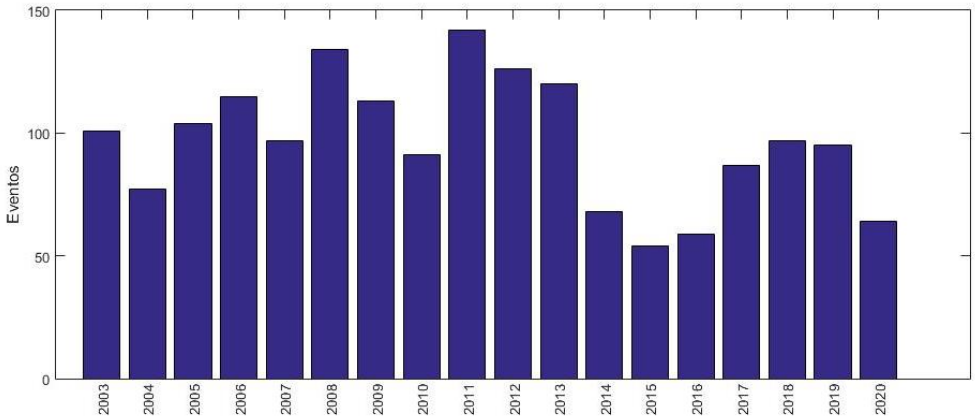


Figura 30. Ocurrencia de eventos FAN a escala mensual durante 2003-2020.

La figura 32 indica la correlación cruzada entre el índice MEI y las anomalías de los eventos FA. En el eje 'x' se indica el desplazamiento de la serie por unidad de tiempo donde se evalúan

un desfase de -20 a +20 meses. En el eje 'y' muestra la correlación de las series según su desplazamiento. Se observa que la correlación más alta es de -0.3740 con un desplazamiento de 5 meses, mientras que la segunda más alta es de -0.3704 sin ningún desplazamiento. La correlación negativa confirma la influencia de eventos interanuales sobre los FA en SAMAO. La relación del fenómeno ENSO con la productividad primaria en CAC ha sido documentada en Germán *et al.,* (2016), donde distinguen oscilaciones en los FA que atribuyen a El Niño, además señalan que la tendencia del volumen y cantidad de FA va en aumento lo que difiere de los resultados encontrados en este trabajo donde, aunque con valores bajos, la tendencia va en decrecimiento. Esta disminución en la tendencia puede deberse al desplazamiento de especies en la comunidad biológica del lago hacia especies más pequeñas no formadoras de natas (*scum*). Los estudios de identificación de especies de fitoplancton en SAMAO indican un cambio en la comunidad que generan cambios en las propiedades ópticas del lago, por lo que se requiere una nueva caracterización y modelación.

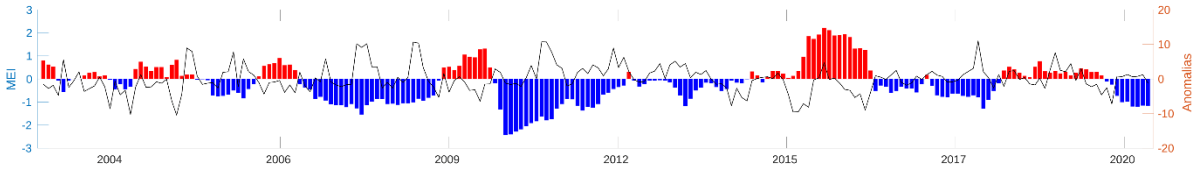


Figura 31. Relación de anomalías con el índice multivariado de El Niño/Oscilación del Sur.

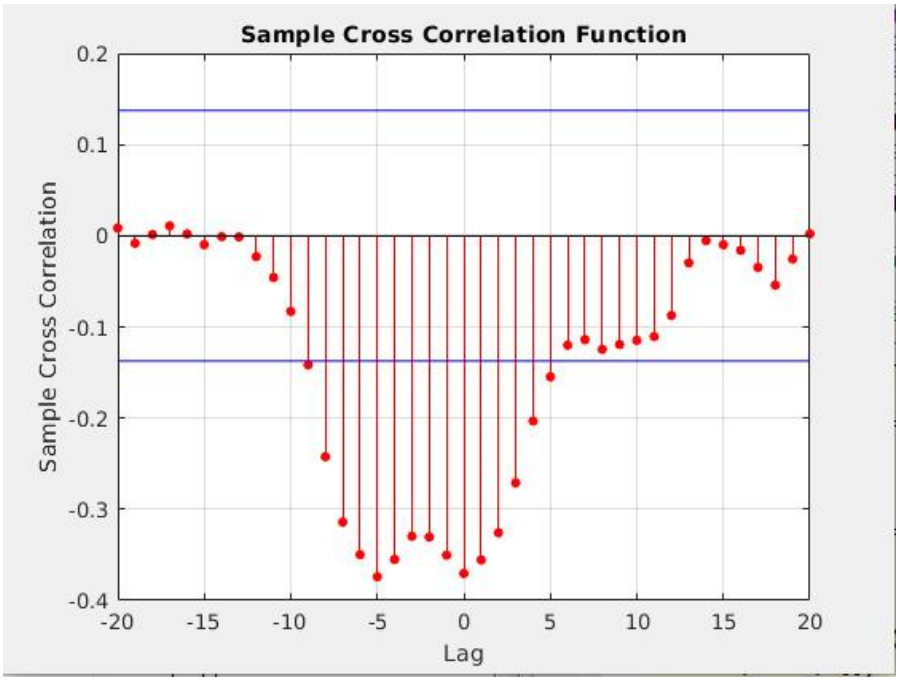


Figura 32. Análisis de correlación cruzada entre el índice MEI y las anomalías de los eventos FA.

Discusión

Las reflectividades en las bandas RED y NIR presentaron comportamiento semejante entre sí, con una dispersión alta en las clases de control y baja en las clases de color del agua, donde, las clases de control tienen valores y dispersión muy altos debido al albedo (Stephens *et al.,* 2015). Por otro lado, como menciona Moses *et al.,* (2017) hay retos al desarrollar algoritmos para CAC, pues el agua es una molécula altamente absorbente y contribuye con solo el 20% de la radiación total que mide el sensor. En este sentido, y de acuerdo a Shi *et al.* (2013) que señalan la importancia de identificar elementos en la BD que aporten mayor información útil para entrenar los algoritmos: los índices clasificatorios propuestos nos permitieron hacer una mejor

caracterización de las clases para entrenamiento al obtener más rasgos propios de éstas, reduciendo así el error en la predicción.

Al evaluar los algoritmos clasificadores RF obtuvo los mejores resultados. En estudios como el de Pal, (2005) y Shaik & Srinivasan, (2019) mencionan particularmente la capacidad de RF como clasificador. Ananias *et al.,* (2022) integraron RF a una cadena automatizada de detección de FA mediante datos MODIS, en la fase de evaluación el algoritmo obtuvo un acierto del 95%, dejando un precedente de la propuesta de este trabajo que es: implementar el algoritmo obtenido en este trabajo para el monitoreo de cuerpos de agua como SAMAO.

En el análisis temporal observamos un patrón estacional donde mayo, abril y marzo son, en ese orden, los meses con la mayor acumulación de eventos, lo que coincide con los resultados encontrados por Cortes-Macias (2018) y Salazar *et al.,* (2021) quiénes en 2015 monitorearon el florecimiento que tuvo lugar de enero a abril. Los meses con menor incidencia de FA encontrados van de julio a enero, cuando el lago permanece estratificado (Sosa-Najéra *et al*., 2010). Germán *et al*., 2016 encontraron un comportamiento periódico en los FA con picos en verano y valles en invierno empleando una BD MODIS del período 2001-2014, por su parte, Shi *et al.,* (2019) evaluaron un modelo NIR-RED con datos de entre 2013-2017 de MODIS y reportaron también ciclos estacionales con picos anuales de FA en julio-agosto.

Encontramos también que los píxeles con mayor incidencia de FA se ubican en la región sureste del lago. El viento es uno de los principales componentes en la distribución espacial superficial de las poblaciones de algas en los cuerpos de agua (Hsiao, 1988). Las características orográficas de SAMAO definen un patrón en las fluctuaciones del viento donde, la brisa se mueve al oeste por la noche, al este por la tarde y su rotación es en sentido antihorario. Alrededor

del mediodía, en las horas aproximadas al paso del sensor, Serrano *et al.,* (2002) encontraron que se forma un giro anticiclónico en la parte sur del lago lo que puede explicar por qué estos píxeles han sido los de mayor incidencia de FA detectados.

Al descomponer la serie completa de FA obtuvimos el ciclo estacional, las anomalías y la tendencia de estos eventos. Se encontró una tendencia negativa que indica que los florecimientos en SAMAO propenden a su reducción. Esto difiere con lo encontrado por Shi *et al.,* 2016 y Germán *et al.,* 2016 en el lago Taihu y el embalse San Roque respectivamente. En el caso de San Roque, se observó que, en los últimos 5 años, la tendencia del volumen y cantidad de FA fue en aumento. Por su parte, Shi *et al.,* 2016 estudiaron los factores involucrados en incremento de FA en el lago Taihu, concluyeron que se debe al aumento de temperatura seguido de las concentraciones de fósforo en el agua. En SAMAO estos factores son de gran importancia también, sin embargo, esta tendencia en decrecimiento puede deberse al desplazamiento de especies en la comunidad biológica del lago, puesto que entre febrero y marzo del 2021 ocurrió en el lago un cambio de color a marrón que indica la presencia de especies diferentes en la superficie del lago.

Al sobreponer las anomalías encontradas en la serie FA obtenida con el índice MEI se encontró una aparente relación del fenómeno ENSO con la biomasa en SAMAO. Resultados similares han sido documentados en German *et al.,* (2016) donde distinguen oscilaciones en los FA de San Roque que atribuyen a El Niño. En México, El Niño impacta el clima provocando mayor precipitación en invierno y escasez de lluvia en verano. Los inviernos con Niño resultan más fríos en casi todo el país, mientras que los veranos son más secos y cálidos. Por su parte, durante los veranos en donde se presenta la Niña las lluvias parecen estar por encima de lo

normal en la mayor parte de México, especialmente en la costa noroeste del Pacífico (Magaña, *et al.*, 1999). En este trabajo se encontraron anomalías positivas de FA que ocurrieron durante períodos la Niña del ENSO y anomalías negativas de FA durante períodos el Niño. Estos resultados pueden atribuirse al incremento de escorrentía por el aumento de lluvia en la región, pero es necesario realizar más estudios que corroboren estos resultados.

Conclusiones

Los productos diarios de MODIS proporcionan series temporales largas con las cuales se logró extraer la señal del ciclo anual y la variabilidad interanual (2003-2020) de los FA. Los productos MOD09GQv006/MYD09GQv006 con resolución de 250 m por píxel permitieron realizar estudios históricos de florecimientos algales en cuerpos de agua continentales con superficies relativamente pequeñas (4 km^2).

El ciclo anual de los florecimientos algales en el lago-cráter de SAMAO presenta su máxima frecuencia entre mayo y un mínimo en octubre, donde la mayor incidencia se observa en la región noreste del lago. Por su parte, el análisis interanual de las anomalías en la frecuencia de florecimientos algales presentó una distribución temporal inversa a las anomalías del índice MEI, lo que indica la influencia del ENSO en la ocurrencia de FA. Se observa una mayor presencia de FA durante los períodos de La Niña, mientras que durante los períodos El Niño hay un descenso en la ocurrencia media de eventos. La tendencia de largo período en la serie temporal indica que los eventos FA en SAMAO propenden a su reducción. Esta reducción pudiera ser indicativo de cambios en la respuesta de la comunidad que sostiene los florecimientos algales en el cuerpo de agua.

Para robustecer estos resultados y corroborar las observaciones de relación entre ambos fenómenos es necesario continuar monitoreando las concentraciones de *Chl-a* en SAMAO, así como obtener mediciones de temperatura y precipitación *in-situ*.

Referencias

Airs, R. L., Temperton, B., Sambles, C., Farnham, G., Skill, S. C., & Llewellyn, C. A. (2014). Chlorophyll f and chlorophyll d are produced in the cyanobacterium Chlorogloeopsis fritschii when cultured under natural light and near-infrared radiation. Febs Letters, 588(20), 3770-3777.

Alcântara, E., De Andrade, C. P., Gomes, A. C., Bernardo, N., Carmo, A. F., Rodrigues, T., & Watanabe, F. (2018, July). Performance analysis of the c2rcc processor in estimate the water quality parameters in inland waters using olci/sentinel-3a images. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium (pp. 9300-9303). Ieee.

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Armienta, M. A., Vilaclara, G., De la Cruz-Reyna, S., Ramos, S., Ceniceros, N., Cruz, O., ... & Arcega-Cabrera, F. (2008). Water chemistry of lakes related to active and inactive Mexican volcanoes. Journal of Volcanology and Geothermal Research, 178(2), 249-258.

Arreola, Karla Susana Barrón, and María Alicia Fonseca Morales. n.d. “Temas Selectos de Turismo y Sustentabilidad.”

Ayeni, A. O., & Adesalu, T. A. (2018). Validating chlorophyll-a concentrations in the Lagos Lagoon using remote sensing extraction and laboratory fluorometric methods. MethodsX, 5, 1204-1212.

Azevedo, S. M., Carmichael, W. W., Jochimsen, E. M., Rinehart, K. L., Lau, S., Shaw, G. R., & Eaglesham, G. K. (2002). Human intoxication by microcystins during renal dialysis treatment in Caruaru—Brazil. Toxicology, 181, 441-446.

Babani, L., Jadhav, S., & Chaudhari, B. (2016). Scaled conjugate gradient based adaptive ANN control for SVM-DTC induction motor drive. In Artificial Intelligence Applications and Innovations: 12th IFIP WG 12.5 International Conference and Workshops, AIAI 2016, Thessaloniki, Greece, September 16-18, 2016, Proceedings 12 (pp. 384-395). Springer International Publishing.

Baret, F., & Buis, S. (2008). Estimating canopy characteristics from remote sensing observations: Review of methods and associated problems. Advances in land remote sensing: System, modeling, inversion and application, 173-201.

Bartram, J. (Ed.). (2015). Routledge handbook of water and health. Routledge.

Bartram, J., & Ballance, R. (Eds.). (1996). Water quality monitoring: a practical guide to the design and implementation of freshwater quality studies and monitoring programmes. CRC Press.

Bartram, J., Carmichael, W. W., Chorus, I., Jones, G., & Skulberg, O. (1999). Eu trophication, cyanobacterial blooms and surface scums. Toxic cyanobacteria in water. E & FN Spon, London, UK, 5-7.

Batra, M., & Agrawal, R. (2018). Comparative analysis of decision tree algorithms. In Nature Inspired Computing: Proceedings of CSI 2015 (pp. 31-36). Springer Singapore.

Behrenfeld, M. J., & Falkowski, P. G. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentration. Limnology and oceanography, 42(1), 1-20.

Blix, K., Li, J., Massicotte, P., & Matsuoka, A. (2019). Developing a new machine-learning algorithm for estimating chlorophyll-a concentration in optically complex waters: A case study for high northern latitude waters by using Sentinel 3 OLCI. Remote Sensing, 11(18), 2076.

Breiman, Leo. (1996). "Bagging Predictors." *Machine Learning* 24 (2): 123–40.

Breiman, Leo (2001). "Random Forests." *Machine Learning* 45: 5–32.

Brezonik, P., Menken, K. D., & Bauer, M. (2005). Landsat-based remote sensing of lake water quality characteristics, including chlorophyll and colored dissolved organic matter (CDOM). Lake and Reservoir Management, 21(4), 373-382.

Bro, R., & Smilde, A. K. (2014). Principal component analysis. Analytical methods, 6(9), 2812-2831.

Brockmann, C., Doerffer, R., Peters, M., Kerstin, S., Embacher, S., & Ruescas, A. (2016, August). Evolution of the C2RCC neural network for Sentinel 2 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters. In Living Planet Symposium (Vol. 740, p. 54).

Caicedo, J. P. R., Verrelst, J., Muñoz-Marí, J., Moreno, J., & Camps-Valls, G. (2014). Toward a semiautomatic machine learning retrieval of biophysical parameters. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7(4), 1249-1259.

Camps-Valls, G., Gómez-Chova, L., Muñoz-Marí, J., Vila-francés, J., Amorós-López, J., & Calpe-Maravilla, J. (2006). Retrieval of oceanic chlorophyll concentration with relevance vector machines. Remote Sensing of Environment, 105(1), 23-33.

Cantoral Uriza, Enrique Arturo, Antonia Dolores Asencio Martínez, and Marina Aboal Sanjurjo. "Cianotoxinas: efectos ambientales y sanitarios. Medidas de prevención." Hidrobiológica 27.2 (2017): 241-251.

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., & Xue, K. (2020). A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes. Remote Sensing of Environment, 248, 111974.

Carlson, R. E. (1977). A trophic state index for lakes 1. Limnology and oceanography, 22 (2), 361–369.

Carpenter, S. R. (2005). Eutrophication of aquatic ecosystems: bistability and soil phosphorus. Proceedings of the National Academy of Sciences, 102(29), 10002-10005.

Carpenter, S. R., Stanley, E. H., & Vander Zanden, M. J. (2011). State of the world's freshwater ecosystems: physical, chemical, and biological changes. Annual review of Environment and Resources, 36, 75-99.

Chegoonian, A. M., Zolfaghari, K., Baulch, H. M., & Duguay, C. R. (2021, July). Support vector regression for chlorophyll-a estimation using Sentinel-2 images in small waterbodies. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS (pp. 7449-7452). IEEE.

Chen, M., Li, Y., Birch, D., & Willows, R. D. (2012). A cyanobacterium that contains chlorophyll f–a red-absorbing photopigment. FEBS letters, 586(19), 3249-3254..

Comiso, J. C., McClain, C. R., Sullivan, C. W., Ryan, J. P., & Leonard, C. L. (1993). Coastal Zone Color Scanner pigment concentrations in the Southern Ocean and relationships to geophysical surface features. Journal of Geophysical Research: Oceans, 98(C2), 2419-2451.

Congalton, Russell G. (1991). “A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data.” *Remote Sensing of Environment* 37 (1): 35–46.

Cooper, Gregory F, and Edward Herskovits. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* 9 (4): 309–47.

Cortes-Macías, L. Z. (2018). Validación y calibración del algoritmo OC2 para Landsat 8 aplicado al lago-cráter de Santa María del Oro. Xalisco, Nayarit, México. Unidad Académica de Agricultura, Universidad Autónoma de Nayarit.

Craig, S. E., Jones, C. T., Li, W. K., Lazin, G., Horne, E., Caverhill, C., y Cullen, J. J. (2012). Deriving optical metrics of coastal phytoplankton biomass from ocean colour. Remote Sensing of Environment, 119, 72–83

Cui, T. W., Zhang, J., Wang, K., Wei, J. W., Mu, B., Ma, Y., ... & Chen, X. Y. (2020). Remote sensing of chlorophyll a concentration in turbid coastal waters based on a global optical water classification system. ISPRS Journal of Photogrammetry and Remote Sensing, 163, 187-201.

Dall'Olmo, G., Gitelson, A. A., Rundquist, D. C., Leavitt, B., Barrow, T., y Holz, J. C. (2005). Assessing the potential of seawifs and modis for estimating chlorophyll concentration in turbid productive waters using red and near-infrared bands. Remote Sensing of Environment, 96 (2), 176–187.

De Jong, Sijmen. (1993). SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems* 18 (3): 251–63.

De Ville, Barry. (2013). Decision Trees. *Wiley Interdisciplinary Reviews: Computational Statistics* 5 (6): 448–55.

Debnath, L., & Mikusinski, P. (2005). Introduction to Hilbert spaces with applications. Academic press.

DeFries, R. S., & Chan, J. C. W. (2000). Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. Remote Sensing of Environment, 74(3), 503-515.

Dekker, A., & Peters, S. (1993). The use of the thematic mapper for the analysis of eutrophic lakes: a case study in the netherlands. International Journal of Remote Sensing, 14 (5), 799–821.

Dekker, A. G., Zamurović-Nenad, Ž., Hoogenboom, H. J., & Peters, S. W. M. (1996). Remote sensing, ecological water quality modelling and in situ measurements: a case study in shallow lakes. Hydrological sciences journal, 41(4), 531-547.

Doerffer, Roland, & Helmut Schiller. (2007). The MERIS Case 2 Water Algorithm. *International Journal of Remote Sensing* 28 (3-4): 517–35.

Dörnhöfer, Katja, & Natascha Oppelt. (2016). Remote Sensing for Lake Research and Monitoring–Recent Advances. *Ecological Indicators* 64: 105–22.

Duan, H., Ma, R., & Hu, C. (2012). Evaluation of remote sensing algorithms for cyanobacterial pigment retrievals during spring bloom formation in several lakes of east China. Remote Sensing of Environment, 126, 126–135

Duan, H., Ma, R., Zhang, Y., Loiselle, S. A., Xu, J., Zhao, C., ... & Shang, L. (2010). A new three-band algorithm for estimating chlorophyll concentrations in turbid inland lakes. Environmental Research Letters, 5(4), 044009.

Duan, H., Zhang, Y., Zhang, B., Song, K., & Wang, Z. (2007). Assessment of chlorophyll-a concentration and trophic state for Lake Chagan using Landsat TM and field spectral data. Environmental monitoring and assessment, 129, 295-308.

Dube, T., Mutanga, O., Seutloali, K., Adelabu, S., & Shoko, C. (2015). Water quality monitoring in sub-Saharan African lakes: a review of remote sensing applications. African Journal of Aquatic Science, 40(1), 1-7.

Duc-Hung, L., Cong-Kha, P., Trang, N. T. T., & Tu, B. T. (2012, August). Parameter extraction and optimization using Levenberg-Marquardt algorithm. In 2012 Fourth International Conference on Communications and Electronics (ICCE) (pp. 434-437). IEEE.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z. I., Knowler, D. J., Lévêque, C., ... & Sullivan, C. A. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. Biological reviews, 81(2), 163-182.

Efron, Bradley. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM.

Efron, Bradley (2013). Bayes' Theorem in the 21st Century. *Science* 340 (6137): 1177–78.

Feng, L., Hu, C., Han, X., Chen, X., y Qi, L. (2014). Long-term distribution patterns of chlorophyll-a concentration in china's largest freshwater lake: Meris full-resolution observations with a practical approach. remote sensing, 7 (1), 275–299.

Filazzola, A., Mahdiyan, O., Shuvo, A., Ewins, C., Moslenko, L., Sadid, T., ... & Sharma, S. (2020). A database of chlorophyll and water chemistry in freshwater lakes. Scientific data, 7(1), 1-10.

Fisher, Ronald A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 7 (2): 179–88.

Fisheries, FAO. (2011). Aquaculture Department. 2013. *Global Aquaculture Production Statistics for the Year*.

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), 238-247.

Flach, Peter A, and Nicolas Lachiche. (2004). Naive Bayesian Classification of Structured Data. *Machine Learning* 57 (3): 233–69.

German, Alba, Verónica Andreo, Carolina Tauro, C Marcelo Scavuzzo, and Anabella Ferral. 2020. "A Novel Method Based on Time Series Satellite Data Analysis to Detect Algal Blooms." *Ecological Informatics* 59: 101131.

Germán, A., Tauro, C., Andreo, V., Bernasconi, I., & Ferral, A. (2016, June). Análisis de una serie temporal de clorofila-a a partir de imágenes MODIS de un embalse eutrófico. In 2016 IEEE Biennial Congress of Argentina (ARGENCON) (pp. 1-6). IEEE.

German, A., Andreo, V., Tauro, C., Scavuzzo, C. M., & Ferral, A. (2020). A novel method based on time series satellite data analysis to detect algal blooms. Ecological Informatics, 59, 101131.

Gitelson, Anatoly. (1992). The Peak Near 700 Nm on Radiance Spectra of Algae and Water: Relationships of Its Magnitude and Position with Chlorophyll Concentration. *International Journal of Remote Sensing* 13 (17): 3367–73.

Gitelson, A. A., Dall'Olmo, G., Moses, W., Rundquist, D. C., Barrow, T., Fisher, T. R., ... & Holz, J. (2008). A simple semi-analytical model for remote estimation of chlorophyll-a in turbid waters: Validation. Remote Sensing of Environment, 112(9), 3582-3593.

Gitelson, A., Gurlin, D., Moses, W., Rundquist, D., Leavitt, B., & Barrow, T. (2009). Estimating chlorophyll-a concentration in inland, estuarine and coastal waters: from close range to satellite observations. Advances in Environmental Remote Sensing: Sensors, Algorithms, and Applications.

Gleick, Peter H. (1996). Basic Water Requirements for Human Activities: Meeting Basic Needs. *Water International* 21 (2): 83–92.

Glibert, Patricia M. (2020). Harmful Algae at the Complex Nexus of Eutrophication and Climate Change. *Harmful Algae* 91: 101583.

Gons, Herman J. (1999). Optical Teledetection of Chlorophyll a in Turbid Inland Waters. *Environmental Science & Technology* 33 (7): 1127–32.

Gons, H. J., Auer, M. T., y Effler, S. W. (2008). Meris satellite chlorophyll mapping of oligotrophic and eutrophic waters in the laurentian great lakes. Remote Sensing of Environment, 112 (11), 4098–4106.

Gossn, J. I., Ruddick, K. G., & Dogliotti, A. I. (2019). Atmospheric correction of OLCI imagery over extremely turbid waters based on the red, NIR and 1016 nm bands and a new baseline residual technique. Remote Sensing, 11(3), 220.

Guan, Xian, Jonathan Li, & William G Booty. (2011). Monitoring Lake Simcoe Water Clarity Using Landsat-5 TM Images. *Water Resources Management* 25 (8): 2015–33.

Han, L., & Jordan, K. J. (2005). Estimating and mapping chlorophyll-a concentration in Pensacola Bay, Florida using Landsat ETM+ data. International Journal of Remote Sensing, 26(23), 5245-5254.

Harper, D. M. (1992). Eutrophication of freshwaters (p. 327). London: Chapman & Hall.

Hernández-Morales, R., Hidalgo-Anguiano, M., Murillo, M. D. R. O., & Ríos, M. S. A. (2014). Factores abióticos que rigen la presencia y permanencia del género Microcystis Kützing ex Lemmermann en un lago tropical profundo. Biológicas Revista de la DES Ciencias Biológico Agropecuarias Universidad Michoacana de San Nicolás de Hidalgo, 16(1), 33-42.

Hoerl, Arthur E, & Robert W Kennard. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12 (1): 55–67.

Hooker, Stanford Baird. (1992). SeaWiFS Technical Report Series: An Overview of SeaWiFS and Ocean Color.

Hovis, W. A. (1980). DK Clark, F. Anderson, RW Austin, WH Wilson, ET Baker. D. Ball, HR Gordon, JL Muller, SZ El-Sayed, B. Sturm, RC Wrigley, and CS. Yentsch, Nimbus-7 Coastal Zone Color Scanner: System description and initial imagery, Science, 210(3), 60-63.

Hsiao, S. I. (1988). Spatial and seasonal variations in primary production of sea ice microalgae and phytoplankton in Frobisher Bay, Arctic Canada. Marine Ecology Progress Series, 275-285.

Huang, C., Zou, J., Li, Y., Yang, H., Shi, K., Li, J., ... & Zheng, F. (2014). Assessment of NIR-red algorithms for observation of chlorophyll-a in highly turbid inland waters in China. ISPRS journal of photogrammetry and remote sensing, 93, 29-39.

Hutchinson, G. E., & Eutrophication, P. (1969). Present. Eutrophication: Causes, Consequences, Correctives, Nat. Acad. Sci. Washington, 17-28.

Image Processing Lab (IPL). (2016). "Simple-r: A Simple MATLAB Regression Toolbox." *https://github.com/IPL-UV/simpleR*.

Jain, Anil K. (2010). Data Clustering: 50 Years Beyond k-Means. *Pattern Recognition Letters* 31 (8): 651–66.

Jain, Anil K, Jianchang Mao, & K Moidin Mohiuddin. (1996). Artificial Neural Networks: A Tutorial. *Computer* 29 (3): 31–44.

Jeong, B., Chapeta, M. R., Kim, M., Kim, J., Shin, J., & Cha, Y. (2022). Machine learning-based prediction of harmful algal blooms in water supply reservoirs. Water Quality Research Journal, 57(4), 304-318.

Jia, T., Zhang, X., & Dong, R. (2019). Long-term spatial and temporal monitoring of cyanobacteria blooms using MODIS on google earth engine: A case study in Taihu Lake. Remote Sensing, 11(19), 2269.

Jiang, G., Loiselle, S. A., Yang, D., Ma, R., Su, W., & Gao, C. (2020). Remote estimation of chlorophyll a concentrations over a wide range of optical conditions based on water classification from VIIRS observations. Remote Sensing of Environment, 241, 111735.

Keller, S., Maier, P. M., Riese, F. M., Norra, S., Holbach, A., Börsig, N., ... & Hinz, S. (2018). Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity. International journal of environmental research and public health, 15(9), 1881.

Koduvely, Hari M. (2015). *Learning Bayesian Models with r*. Packt Publishing Ltd.

Kotchenova, SY, & EF Vermote. (2006). A Vector Version of the 6s Radiative Transfer Code for Atmospheric Correction of Satellite Data. In *AGU Fall Meeting Abstracts*, 2006: A13B–0887

Kravitz, J., Matthews, M., Bernard, S., & Griffith, D. (2020). Application of Sentinel 3 OLCI for chl-a retrieval over small inland water targets: Successes and challenges. Remote Sensing of Environment, 237, 111562.

Krinner, Gerhard. (2003). Impact of Lakes and Wetlands on Boreal Climate. *Journal of Geophysical Research: Atmospheres* 108 (D16).

Kutser, T., Herlevi, A., Kallio, K., & Arst, H. (2001). A hyperspectral model for interpretation of passive optical remote sensing data from turbid lakes. Science of the Total Environment, 268(1-3), 47-58.

Kutser, T., Metsamaa, L., Strömbeck, N., & Vahtmäe, E. (2006). Monitoring cyanobacterial blooms by satellite remote sensing. Estuarine, Coastal and Shelf Science, 67(1-2), 303-312.

Larkin, PA, TG Northcote, PA Larkin, TG Northcote, & GA Rohlich. (1969). Eutrophication: Causes, Consequences, Correctives. *Nat. Acad. Sci. Publ* 1700: 256–73.

Le, C., Li, Y., Zha, Y., Sun, D., Huang, C., y Lu, H. (2009). A four-band semi-analytical model for estimating chlorophyll a in highly turbid lakes: The case of taihu lake, China. Remote Sensing of Environment, 113 (6), 1175–1182

Le, C., Li, Y., Zha, Y., Sun, D., Huang, C., & Zhang, H. (2011). Remote estimation of chlorophyll a in optically complex waters based on optical classification. Remote Sensing of Environment, 115(2), 725-737.

Lewis Jr, W. M. (1983). Temperature, heat, and mixing in Lake Valencia, Venezuela 1. Limnology and oceanography, 28(2), 273-286.

Li, J., Gao, M., Feng, L., Zhao, H., Shen, Q., Zhang, F., ... & Zhang, B. (2019). Estimation of chlorophyll-a concentrations in a highly turbid eutrophic lake using a classification-based MODIS land-band algorithm. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(10), 3769-3783.

Lin, S., Pierson, D. C., & Mesman, J. P. (2023). Prediction of algal blooms via data-driven machine learning models: an evaluation using data from a well-monitored mesotrophic lake. Geoscientific Model Development, 16(1), 35-46.

Liu, G., Li, L., Song, K., Li, Y., Lyu, H., Wen, Z., ... & Shi, K. (2020). An OLCI-based algorithm for semi-empirically partitioning absorption coefficient and estimating chlorophyll a concentration in various turbid case-2 waters. Remote Sensing of Environment, 239, 111648.

Lundberg, J. G., Kottelat, M., Smith, G. R., Stiassny, M. L., & Gill, A. C. (2000). So many fishes, so little time: an overview of recent ichthyological discovery in continental waters. Annals of the Missouri Botanical Garden, 26-62.

Lynch, A. J., Cooke, S. J., Deines, A. M., Bower, S. D., Bunnell, D. B., Cowx, I. G., ... & Beard Jr, T. D. (2016). The social, economic, and environmental importance of inland fish and fisheries. Environmental Reviews, 24(2), 115-121.

Lyu, H., Li, X., Wang, Y., Jin, Q., Cao, K., Wang, Q., & Li, Y. (2015). Evaluation of chlorophyll-a retrieval algorithms based on MERIS bands for optically varying eutrophic inland lakes. Science of the Total Environment, 530, 373-382.

Magaña, V., PÉREZ, J. L., & CONDE ÁLVAREZ, C. E. C. I. L. I. A. El niño y la oscilación del sur, sus impactos en México. Ciencias, (051).

Masocha, M., Dube, T., Nhiwatiwa, T., & Choruma, D. (2018). Testing utility of Landsat 8 for remote assessment of water quality in two subtropical African reservoirs with contrasting trophic states. Geocarto international, 33(7), 667-680.

Matthews, M. W. (2017). Bio-optical modeling of phytoplankton chlorophyll-a. In Bio-optical modeling and remote sensing of inland waters (pp. 157-188). Elsevier.

Méndez Reyes, L. S. (2018). Determinación de los coeficientes de exportación de nutrientes en la cuenca del lago Lanalhue, Región del BioBío, Chile.

Mishra, S., & Mishra, D. R. (2012). Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. Remote Sensing of Environment, 117, 394-406.

Mobley, CD. (1994). Light and Water: Radiative Transfer in Natural Waters' Academic Press. *San Diego, California*.

Moore, T. S., Campbell, J. W., & Feng, H. (2001). A fuzzy logic classification scheme for selecting and blending satellite ocean color algorithms. IEEE Transactions on Geoscience and Remote sensing, 39(8), 1764-1776.

Moore, T. S., Dowell, M. D., Bradt, S., & Verdu, A. R. (2014). An optical water type framework for selecting and blending retrievals from bio-optical algorithms in lakes and coastal waters. Remote sensing of environment, 143, 97-111.

Moses, W. J., Gitelson, A. A., Berdnikov, S., & Povazhnyy, V. (2009). Satellite estimation of chlorophyll-$ a $ concentration using the red and NIR bands of MERIS—the Azov sea case study. IEEE Geoscience and Remote Sensing Letters, 6(4), 845-849.

Moses, W. J., Sterckx, S., Montes, M. J., De Keukelaere, L., & Knaeps, E. (2017). Atmospheric correction for inland waters. In Bio-optical modeling and remote sensing of inland waters (pp. 69-100). Elsevier.

Moss, Brian. (2012). Cogs in the Endless Machine: Lakes, Climate Change and Nutrient Cycles: A Review. *Science of the Total Environment* 434: 130–42.

Muñoz-Marí, Jordi, and Gustavo Camps-Valls. (2013) (accessed October 21, 2020). *simpleClass: Simple Classification Toolbox*. *https://github.com/IPL-UV/simpleClass*.

Nas, B., Ekercin, S., Karabörk, H., Berktay, A., & Mulla, D. J. (2010). An application of Landsat-5TM image data for water quality mapping in Lake Beysehir, Turkey. Water, Air, & Soil Pollution, 212, 183-197.

O'Connor, B., & Secades, C. (2013). Review of the use of remotely-sensed data for monitoring biodiversity change and tracking progress towards the Aichi Biodiversity Targets.

O'Neil, J. M., Davis, T. W., Burford, M. A., & Gobler, C. J. (2012). The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. Harmful algae, 14, 313-334.

O'Reilly, J. E., & Werdell, P. J. (2019). Chlorophyll algorithms for ocean color sensors-OC4, OC5 & OC6. Remote sensing of environment, 229, 32-47.

Ochoa-Zamora, G. G. 2018. "Variacion Espacio-Temporal de Las Poblaciones de Cianobacterias Formadoras de Florecimientos En El Lago Cráter de Santa María Del Oro, Nayarit.". Tesis de licenciatura.: Unidad Académica de Agricultura, Universidad Autónoma de Nayarit. Xalisco, Nayarit, México.

Ogashawara, Igor. (2019). The Use of Sentinel-3 Imagery to Monitor Cyanobacterial Blooms. *Environments* 6 (6): 60.

Paerl, Hans W, & David F Millie. (1996). Physiological Ecology of Toxic Aquatic Cyanobacteria. *Phycologia* 35 (sup6): 160–67.

Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222.

Pahlevan, N., Smith, B., Schalles, J., Binding, C., Cao, Z., Ma, R., ... & Stumpf, R. (2020). Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. Remote Sensing of Environment, 240, 111604.

Palmer, S. C., Hunter, P. D., Lankester, T., Hubbard, S., Spyrakos, E., Tyler, A. N., ... & Tóth, V. R. (2015). Validation of Envisat MERIS algorithms for chlorophyll retrieval in a large, turbid and optically-complex shallow lake. Remote Sensing of Environment, 157, 158-169.

Pelckmans, K., Suykens, J. A., Van Gestel, T., De Brabanter, J., Lukas, L., Hamers, B., ... & Vandewalle, J. (2002). LS-SVMlab: a matlab/c toolbox for least squares support vector machines. Tutorial. KULeuven-ESAT. Leuven, Belgium, 142(1-2).

Pértegas Díaz, S., & Pita Fernández, S. (2001). La distribución normal. Cad Aten Primaria, 8, 268-274.

Pizzolon, L. I. N. O. (1996). Importancia de las cianobacterias como factor de toxicidad en las aguas continentales. Interciencia, 21(6), 239-245.

Porras, P. (2018). Procesos Gaussianos para problemas de regresión y estimación de la incertidumbre.

Prasad, S., Saluja, R., & Garg, J. (2020). Assessing the efficacy of landsat-8 oli imagery derived models for remotely estimating chlorophyll-a concentration in the upper ganga river, india. International Journal of Remote Sensing, 41 (7), 2439–2456.

Qi, L., Hu, C., Duan, H., Barnes, B. B., & Ma, R. (2014). An eof-based algorithm to estimate chlorophyll a concentrations in Taihu lake from modis land-band measurements: Implications for near real-time applications and forecasting models. Remote Sensing, 6 (11), 10694–10715

Raileanu, Laura Elena, & Kilian Stoffel. 2004. “Theoretical Comparison Between the Gini Index and Information Gain Criteria.” *Annals of Mathematics and Artificial Intelligence* 41 (1): 77–93.

Ramus, J. (1995). Submarine Light and Primary Production–Light and Photosynthesis in Aquatic Ecosystems. By John TO Kirk. *BioScience* 45 (3): 220.

Rani, M., Rehman, S., Sajjad, H., Sidiki Alare, R., Chaudhary, B. S., Patairiya, S., ... & Kumar, P. (2019). NIR-red algorithms-based model for chlorophyll-a retrieval in highly turbid Inland Densu River Basin in South-East Ghana, West Africa. IET Image Processing, 13(8), 1328-1332.

Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning, vol. 1.

Reynolds, C. S., Oliver, R. L., & Walsby, A. E. (1987). Cyanobacterial dominance: the role of buoyancy regulation in dynamic lake environments. New Zealand journal of marine and freshwater research, 21(3), 379-390.

Rosipal, R., & Krämer, N. (2006). Overview and recent advances in partial least squares. In Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop, SLSFS 2005, Bohinj, Slovenia, February 23-25, 2005, Revised Selected Papers (pp. 34-51). Springer Berlin Heidelberg.

Roodschild, M., Gotay Sardiñas, J., Will, A. E., & Rodriguez, S. A. (2019). Optimización de Scaled Conjugate Gradient para Froog Neural Networks. In XX Simposio Argentino de Inteligencia Artificial (ASAI 2019)-JAIIO 48 (Salta).

Ruddick, K. G., Gons, H. J., Rijkeboer, M., & Tilstone, G. (2001). Optical remote sensing of chlorophyll a in case 2 waters by use of an adaptive two-band algorithm with optimal error properties. Applied optics, 40 (21), 3575–3585.

Sagi, Omer, & Lior Rokach. 2018. "Ensemble Learning: A Survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8 (4): e1249.

Sala, O. E., Stuart Chapin, F. I. I. I., Armesto, J. J., Berlow, E., Bloomfield, J., Dirzo, R., ... & Wall, D. H. (2000). Global biodiversity scenarios for the year 2100. science, 287(5459), 1770-1774.

Salas-Betancourt, Alfredo. 2017. "Dinámica de Nutrientes del Lago De Santa María del Oro, Nayarit.". Tesis de licenciatura.: Universidad Autónoma de Nayarit. Nayarit, México

Salazar-Alcaraz, Ivan. 2018. "Identificación y Aislamiento de Cianobacterias de Un Lago Cráter Tropical." Tesis de Maestría, Universidad Autónoma de Nayarit. Xalisco, Nayarit, México.

Samuel, Arthur L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* 3: 210–29.

Samui, P., Bhattacharya, G., & Das, S. K. (2008). Support vector machine and relevance vector machine classifier in analysis of slopes. International association for computer methods and advances in geomechanics (IACMAG), 1-6.

Schalles, J. F. (2006). Optical remote sensing techniques to estimate phytoplankton chlorophyll a concentrations in coastal. In Remote sensing of aquatic coastal ecosystem processes (pp. 27-79). Dordrecht: Springer Netherlands.

Schowengerdt, Robert A. (2006). *Remote Sensing: Models and Methods for Image Processing*. Elsevier.

Serrano, D., Filonov, A., & Tereshchenko, I. (2002). Dynamic response to valley breeze circulation in Santa Maria del Oro, a volcanic lake in Mexico. Geophysical Research Letters, 29(13), 27-1.

Shaik, A. B., & Srinivasan, S. (2019). A brief survey on random forest ensembles in classification model. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2 (pp. 253-260). Springer Singapore.

Shalev-Shwartz, Shai, & Shai Ben-David. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press.

Shawe-Taylor, J., & Cristianini, N. (2004). Kernel methods for pattern analysis. Cambridge university press.

Shi, K., Li, Y., Li, L., Lu, H., Song, K., Liu, Z., ... & Li, Z. (2013). Remote chlorophyll-a estimates for inland waters based on a cluster-based classification. Science of the Total Environment, 444, 1-15.

Shi, K., Zhang, Y., Xu, H., Zhu, G., Qin, B., Huang, C., ... & Lv, H. (2015). Long-term satellite observations of microcystin concentrations in Lake Taihu during cyanobacterial bloom periods. Environmental Science & Technology, 49(11), 6448-6456.

Shi, K., Zhang, Y., Zhang, Y., Li, N., Qin, B., Zhu, G., & Zhou, Y. (2019). Phenology of phytoplankton blooms in a trophic lake observed from long-term MODIS data. Environmental science & technology, 53(5), 2324-2331.

Shi, K., Zhang, Y., Zhang, Y., Qin, B., & Zhu, G. (2020). Understanding the long-term trend of particulate phosphorus in a cyanobacteria-dominated lake using MODIS-Aqua observations. Science of The Total Environment, 737, 139736.

Shi, K., Zhang, Y., Zhou, Y., Liu, X., Zhu, G., Qin, B., & Gao, G. (2017). Long-term MODIS observations of cyanobacterial dynamics in Lake Taihu: Responses to nutrient enrichment and meteorological factors. Scientific reports, 7(1), 40326.

Shi, K., Zhang, Y., Qin, B., & Zhou, B. (2019). Remote sensing of cyanobacterial blooms in inland waters: present knowledge and future challenges. Science Bulletin, 64(20), 1540-1556.

Singh, K., Ghosh, M., Sharma, S. R., & Kumar, P. (2014). Blue–red–NIR model for chlorophyll-a retrieval in hypersaline–alkaline water using Landsat ETM+ sensor. IEEE Journal of selected topics in applied earth observations and remote sensing, 7(8), 3553-3559.

Soomets, T., Uudeberg, K., Jakovels, D., Zagars, M., Reinart, A., Brauns, A., & Kutser, T. (2019). Comparison of lake optical water types derived from Sentinel-2 and Sentinel-3. Remote Sensing, 11(23), 2883.

Soriano-González, J., Urrego, E. P., Sòria-Perpinyà, X., Angelats, E., Alcaraz, C., Delegido, J., ... & Moreno, J. (2022). Towards the combination of C2RCC processors for improving water quality retrieval in inland and coastal areas. Remote Sensing, 14(5), 1124.

Sosa-Nájera, S., Lozano-García, S., Roy, P. D., & Caballero, M. (2010). Registro de sequías históricas en el occidente de México con base en el análisis elemental de sedimentos lacustres: El caso del lago de Santa María del Oro. Boletín de la Sociedad Geológica Mexicana, 62(3), 437-451.

Spyrakos, Evangelos, et al. "Optical types of inland and coastal waters." Limnology and Oceanography 63.2 (2018): 846-870.

Starczewski, A., & Krzyżak, A. (2015). Performance evaluation of the silhouette index. In Artificial Intelligence and Soft Computing: 14th International Conference, ICAISC 2015, Zakopane, Poland, June 14-18, 2015, Proceedings, Part II 14 (pp. 49-58). Springer International Publishing.

Steinberg, C. E., & Hartmann, H. M. (1988). Planktonic bloom-forming Cyanobacteria and the eutrophication of lakes and rivers. Freshwater Biology, 20(2), 279-287.

Stendera, S., Adrian, R., Bonada, N., Cañedo-Argüelles, M., Hugueny, B., Januschke, K., ... & Hering, D. (2012). Drivers and stressors of freshwater biodiversity patterns across different ecosystems and scales: a review. Hydrobiologia, 696, 1-28.

Stephens, G. L., O'Brien, D., Webster, P. J., Pilewski, P., Kato, S., & Li, J. L. (2015). The albedo of Earth. Reviews of geophysics, 53(1), 141-163.

Stiassny, M. L. J., Parenti, L. R., & Johnson, G. D. (1996). Interrelationships of fishes Academic Press. San Diego. 496s.

Su, H., Lu, X., Chen, Z., Zhang, H., Lu, W., & Wu, W. (2021). Estimating coastal chlorophyll-a concentration from time-series OLCI data based on machine learning. Remote Sensing, 13(4), 576.

Suthers, I., Rissik, D., & Richardson, A. (Eds.). (2019). Plankton: A guide to their ecology and monitoring for water quality. CSIRO publishing.

Suykens, J. A., Lukas, L., & Vandewalle, J. (2000, April). Sparse least squares Support Vector Machine classifiers. In ESANN (pp. 37-42).

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. AI communications, 30(2), 169-190.

The MathWorks, Inc. (2010). *Deep Learning Toolbox*. Natick, Massachusetts, United State. *https://www.mathworks.com/help/deeplearning/ref/patternnet.html*.

Tomaselli, L. (2004). The microalgal cell. Handbook of microalgal culture: biotechnology and applied phycology, 1, 3-19.

Toming, K., Kutser, T., Uiboupin, R., Arikas, A., Vahter, K., & Paavel, B. (2017). Mapping water quality parameters with sentinel-3 ocean and land colour instrument imagery in the Baltic Sea. Remote Sensing, 9(10), 1070.

Torbick, N., Hession, S., Hagen, S., Wiangwang, N., Becker, B., & Qi, J. (2013). Mapping inland lake water quality across the Lower Peninsula of Michigan using Landsat TM imagery. International journal of remote sensing, 34(21), 7607-7624.

Torbick, N., Hu, F., Zhang, J., Qi, J., Zhang, H., & Becker, B. (2008). Mapping chlorophyll-a concentrations in West Lake, China using Landsat 7 ETM+. Journal of Great Lakes Research, 34(3), 559-565.

Tranvik, L. J., Downing, J. A., Cotner, J. B., Loiselle, S. A., Striegl, R. G., Ballatore, T. J., ... & Weyhenmeyer, G. A. (2009). Lakes and reservoirs as regulators of carbon cycling and climate. Limnology and oceanography, 54(6part2), 2298-2314.

Uudeberg, K., Ansko, I., Põru, G., Ansper, A., & Reinart, A. (2019). Using optical water types to monitor changes in optically complex inland and coastal waters. Remote Sensing, 11(19), 2297.

Valdéz Blanco, D. (2010). Regresión por mínimos cuadrados parciales. Revista Varianza, 18.

Valle Moreno, J., & Guerra Bustillo, W. (2012). La multicolinealidad en modelos de regresión lineal múltiple. Revista Ciencias Técnicas Agropecuarias, 21(4), 80-83.

Vapnik, V., Golowich, S., & Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. Advances in neural information processing systems, 9.

Verhoef, W. (1996). Application of harmonic analysis of NDVI time series (HANTS). Fourier analysis of temporal NDVI in the Southern African and American continents, 108, 19-24.

Vermote, E. (2015). MOD09A1 MODIS/terra surface reflectance 8-day L3 global 500m SIN grid V006. NASA EOSDIS Land Processes DAAC, 10.

Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J. P., Veroustraete, F., Clevers, J. G., & Moreno, J. (2015). Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties–A review. ISPRS Journal of Photogrammetry and Remote Sensing, 108, 273-290.

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J. P., Camps-Valls, G., & Moreno, J. (2012). Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and-3. Remote Sensing of Environment, 118, 127-139.

Verrelst, J, and JP Rivera. (2019) (accessed May 11, 2022). "Classification Toolbox." *https://www.artmotoolbox.com/classification-toolbox.html*.

Verrelst, J., Rivera, J., Alonso, L., & Moreno, J. (2011, April). ARTMO: An Automated Radiative Transfer Models Operator toolbox for automated retrieval of biophysical parameters through model inversion. In Proceedings of the EARSeL 7th SIG-Imaging Spectroscopy Workshop, Edinburgh, UK (pp. 11-13).

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2020). A comprehensive survey of loss functions in machine learning. Annals of Data Science, 1-26.

Wang, S., Li, J., Zhang, B., Spyrakos, E., Tyler, A. N., Shen, Q., ... & Peng, D. (2018). Trophic state assessment of global inland waters using a MODIS-derived Forel-Ule index. Remote sensing of environment, 217, 444-460.

Williamson, D. F., Parker, R. A., & Kendrick, J. S. (1989). The box plot: a simple visual method to interpret data. Annals of internal medicine, 110(11), 916-921.

Xing, X. G., Zhao, D. Z., Liu, Y. G., Yang, J. H., Xiu, P., & Wang, L. (2007). An overview of remote sensing of chlorophyll fluorescence. Ocean Science Journal, 42, 49-59.

Zhang, F., Li, J., Shen, Q., Zhang, B., Tian, L., Ye, H., ... & Lu, Z. (2019). A soft-classification-based chlorophyll-a estimation method using MERIS data in the highly turbid and eutrophic Taihu Lake. International Journal of Applied Earth Observation and Geoinformation, 74, 138-149.

Zurawell, R. W. (2015). Toxic cyanobacteria. In Routledge Handbook of Water and Health (pp. 98-106). Routledge.