CIENCIA Y TECNOLOGÍA

Investigaciones Latinoamericanas en Ingeniería y Arquitectura

Revista ILIA N°2, 2025, pp: 5-12
URL: https://revistas.uca.edu.sv/index.php/ilia/index
ISSN: 3078-2740 / e-ISSN: 3078-2759 / ISSN-L: 3078-2740

Modelos estadísticos para estimar la clorofila-A con sentinel-2A del Embalse Cerrón Grande

Statistical models for estimating chlorophyll-A with Sentinel-2A from the Cerrón Grande Reservoir

DOI: https://doi.org/10.51378/ilia.vi2.9665

F. Carranza[1], J. Lemus[1] ,A. Chávez[2]

[1] Maestro en Estadística Aplicada a la Investigación, Universidad Centroamericana José Simeón Cañas, UCA, El Salvador.

[2] Maestra en Estadística Aplicada a la Investigación, Universidad Centroamericana José Simeón Cañas, UCA, El Salvador /Departamento de Matemática, Universidad Centroamericana José Simeón Cañas, UCA, El Salvador.

E-mail: fcarranza@protonmail.com, jlemus@pedagogica.edu.sv y achavez@uca.edu.sv

Fecha de recepción: 16 de mayo de 2025 Fecha de aprobación: 16 de septiembre de 2025

Resumen

Este estudio surge debido a la crisis de contaminación ambiental en el Humedal del Cerrón Grande, que afecta la calidad ecológica del embalse, como también los servicios ecosistémicos que el humedal puede ofrecer. Una de las problemáticas comunes de contaminación que enfrentan los humedales es la eutrofización de sus aguas, generando proliferaciones de fitoplancton que pueden alcanzar niveles peligrosos. La concentración de clorofila-a es una medida indirecta de la biomasa de fitoplancton en un cuerpo de agua [1]. Se tuvo como objetivo estimar el valor de la concentración de clorofila-a mediante modelos empíricos; para ello, se utilizaron bandas del satélite Sentinel-2A que dieran una alternativa al índice más utilizado llamado TBDO (Triple banda Dall' Olmo). Se generaron índices con todas las posibles combinaciones de las bandas; además, mediante el análisis de componentes principales, se agruparon las bandas y la clorofila-a en tres factores que resultaron similares a las bandas utilizadas por el TBDO. Como resultado se eligieron 3 modelos de 100,000 posibilidades generadas, que al compararlos contra el TBDO ofrecen mejor estimación de la concentración de clorofila-a en el Humedal. Se superaron la mayoría de los supuestos de los modelos y se analizó la autocorrelación espacial.

Palabras clave — Clorofila-A, teledetección, estadística, contaminación ambiental, TBDO

Abstract

This study arises due to the environmental pollution crisis in the El Cerrón Grande Wetland, which affects the ecological quality of the reservoir, as well as the ecosystem services that the wetland can offer. One of the common pollution problems faced by wetlands is the eutrophication of their waters, generating proliferations of phytoplankton that can reach dangerous levels. Chlorophyll-a concentration is an indirect measure of the phytoplankton biomass in a body of water [1]. The objective was to estimate the value of the concentration of chlorophyll-a through empirical models; To do this, bands from the Sentinel-2A satellite were used to provide an alternative to the most widely used index called TBDO (Triple Band Dall' Olmo). Indices were generated with all possible combinations of the bands; furthermore, by means of principal component analysis, the bands and chlorophyll-a were grouped into three factors that were similar to the bands used by the TBDO. As a result, 3 models of 100,000 generated possibilities were chosen, which when compared against the TBDO offer a better estimate of the concentration of chlorophyll-a in the Wetland. Most of the assumptions of the models were overcome and the spatial autocorrelation was analyzed.

Keywords — Chlorophyll-A, Remote Sensing, Statistics, Environmental pollution, TBDO

I. INTRODUCCIÓN

En El Salvador, el Embalse Cerrón Grande es uno de los ecosistemas acuáticos más amenazados por la contaminación ambiental, afectando a muchas comunidades que hacen uso del agua de este lugar y a la salud del ecosistema. Es importante destacar que los humedales son áreas importantes de biodiversidad a nivel mundial [2].

Las cianobacterias son un tipo de fitoplancton, y se ha detectado que es la especie predominante en el humedal en la época lluviosa [3]. Los fitoplánctones son los microorganismos base de la cadena alimenticia y fuente de oxígeno en el ecosistema acuífero, pero cuando estos sufren de afloramiento (fenómeno Bloom), el oxígeno puede agotarse con la descomposición del fitoplancton muerto, impactando en la vida acuática del humedal, al morir pezhzces y otros organismos [4].

Partiendo de la importancia de medir el valor de concentración de clorofila-a, como medida indirecta de la biomasa del fitoplancton [4]. El objetivo de esta investigación es generar modelos empíricos basados en los índices de las bandas del Sentinel-2A, comparar los modelos empíricos con los resultados obtenidos por el algoritmo TBDO y comparar los modelos utilizando el análisis estadístico espacial.

Para la obtención de dichos modelos, se utiliza como insumo una muestra del humedal, la cual es parte de los resultados de la investigación que está siendo realizada por un equipo de especialistas en ciencias ambientales y teledetección de la Universidad Centroamericana "José Simeón Cañas"[3], quienes bajo criterios técnicos seleccionaron el satélite Sentinel-2A, utilizaron un muestreo probabilístico estratificado de mínima varianza para la obtención de la concentración de clorofila-a bajo una resolución de 20 metros cuadrados. La muestra proporcionada contiene el resultado de los análisis del laboratorio y las fotografías del humedal con sus respectivos valores digitales provenientes de cada una de las 13 bandas del satélite Sentinel-2A. El total de la población de la muestra fue de 32 puntos georreferenciados.

Se evalúa la necesidad de usar un modelo diferente para la época lluviosa, ya que el grupo más abundante de la comunidad fitoplanctónica es el de las cianobacterias que constituyen en un 99 % la abundancia promedio durante abril, mayo y junio [3].

Las disciplinas involucradas que dan soporte a los modelos resultantes y la aplicación de estos se muestran en la figura 1.

En la metodología se encuentran cada uno de los procesos involucrados, iniciando desde la ingesta de los datos hasta la generación de los modelos resultantes.

Se utilizan 3 tipos de software: SPSS, R y QGIS por lo que a través de los diferentes capítulos se observan técnicas y gráficos provenientes de la especialidad de cada uno de estos programas.

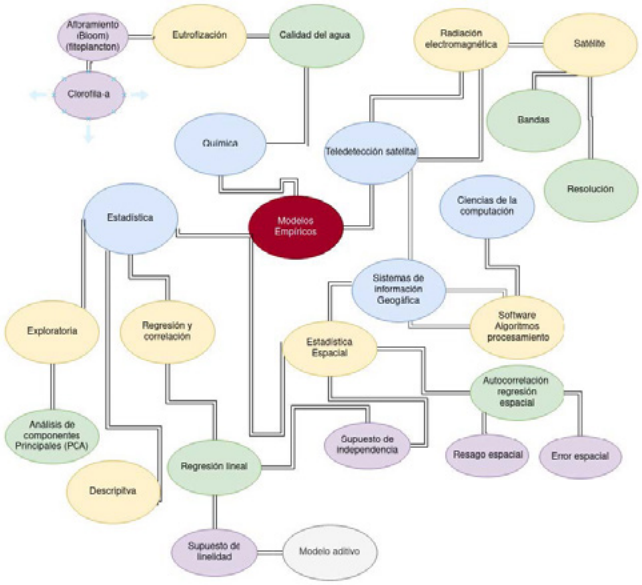

Fig. 1. Conceptos y teoría en los que se basa esta investigación.

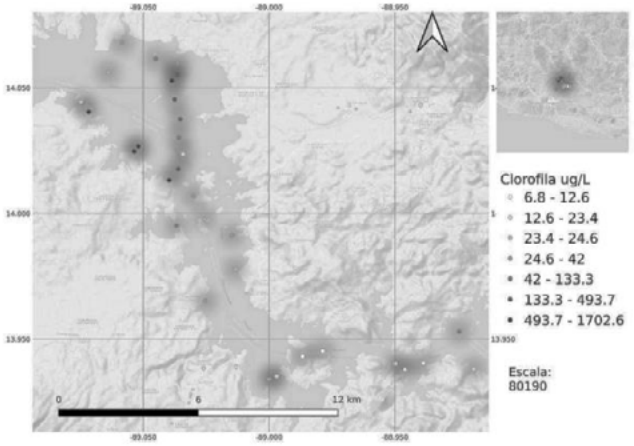

Fig. 2. Visualización de la muestra del Cerrón Grande en un mapa de calor.

Población: concentración de la clorofila-a en el Humedal del Cerrón Grande siendo un total de 226,545 objetos y cada objeto tiene un área 20x20 metros cuadrados.

Muestra: promedio de la concentración de clorofila-a de dos tomas obtenidas en cada uno de 32 puntos georreferenciados del Cerrón Grande y la reflectancia de las 13 bandas del satélite Sentinel-2A de cada punto, se obtuvo mediante muestreo estratificado por mínima varianza.

II. METODOLOGÍA

Se realizó la prueba de Kruskal Wallis con la variable clorofila-a, para saber si los datos provenían de la misma población (figura 3).

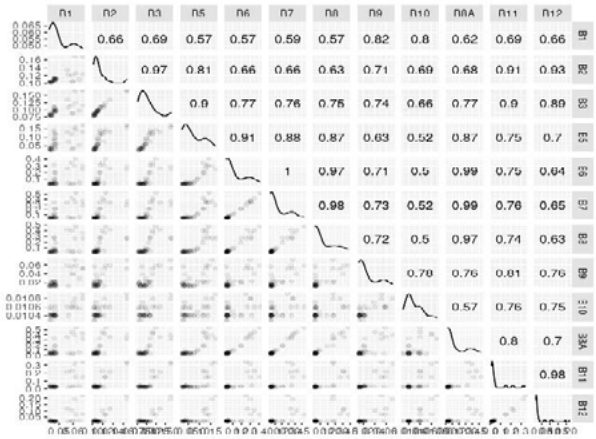

Fig. 3. Diagrama de pares entre las variables de estudio.

Los valores del estrato 4, contiene los valores más bajo de clorofila-a, el estrato 2, tiene la mayor variabilidad y los datos con mayor valor de clorofila-a, en la identificación de los estratos con respecto a su posición en el ejes de las "y" se puede observar que el estrato 1 es el que varía más con respecto a la latitud, para el caso de la posición en el eje de las "x", se puede observar que es el estrato 4 es el que tiene mayor variabilidad, refiriéndose a la zona sur, en los cuadros comparativos de las medidas de laboratorio contra latitud se puede observar que lo valores más altos están ubicados entre el estrato 2 y el estrato 3 para la latitud entre 14.04 y 14.00, para el caso de la longitud y los valores más altos de la clorofila-a estos se ubican entre el -89.04 y -89.07.

Al analizar la correlación lineal (figura 3) entre las bandas y el valor de la clorofila-a se observa que las bandas 5, 6, 7, 8, 8A tienen los mayores valores. De manera exploratoria, usando el método de selección de variables hacia adelante y hacia atrás en la regresión lineal múltiple, el modelo más consistente fue con la banda A8 con R^2 de 0.574. Si bien este resultado es comparable con el algoritmo TBDO, no supera los valores obtenidos con los demás modelos, esto refuerza la relevancia de la banda A8, que es usado en el modelo polinomial presentado más adelante, pero cabe destacar que en los valores menores a 400 ug/l las bandas no logran correlacionar linealmente, se identificó que los estratos 1 y 4 son los que menos tiene correlación lineal con el valor de la clorofila-a, dichos hallazgos son importantes para las siguientes fases de la investigación.

Con el método de rotación Varimax (tabla 1) se han obtenido los siguientes factores:

1. Borde rojo de vegetación, infrarrojo cercano y clorofila-a: B5, B6, B7, B8 y B8A.

2. Azul, verde, rojo e infrarrojo de onda corta: B2, B3, B4, B11 y B12.

3. Aerosol costero, vapor de agua y nubes de cirrus: B1, B10 y B9.

Tabla 1.: Matriz de las 3 componentes con rotación Varimax

Bandas y clorofila-a	Componente 1	Componente 2	Componente 3
B1	0.311	0.294	0.830
B2	0.335	0.885	0.300
B3	0.486	0.814	0.265
B4	0.244	0.920	0.283
B5	0.750	0.574	0.132
B6	0.906	0.342	0.217
B7	0.902	0.325	0.254
B8	0.905	0.320	0.217
B9	0.416	0.358	0.735
B8A	0.874	0.356	0.299
B10	0.160	0.335	0.869
B11	0.452	0.754	0.400
B12	0.316	0.836	0.380
Clorofila-a	0.706	0.154	0.423

En la figura 4 se muestra el flujo de generación de los modelos, en primer lugar, se creó una lista de índices candidatos a ser probados de una manera más sistemática que la forma intuitiva utilizada en la fase preliminar, estos surgieron de operaciones matemáticas entre las bandas con base al estado del arte de índices empíricos, luego se crearon los listados de los modelos a probar sacando provecho de la variedad de modelos que se pueden generar mediante software R.

El procesamiento consistió en elegir cada uno de los índices y para cada uno de ellos se crearon todas las combinaciones posibles entre las 13 bandas, se procedió a generar los modelos por cada combinatoria y se guardaron en diccionarios CSV los resultados: el nombre del modelo, el R^2, R^2 ajustado, AIC, RSE. En una fase posterior se filtraron los modelos con mayor R^2 y menor AIC (ofrece la mejor capacidad de predicción futura).

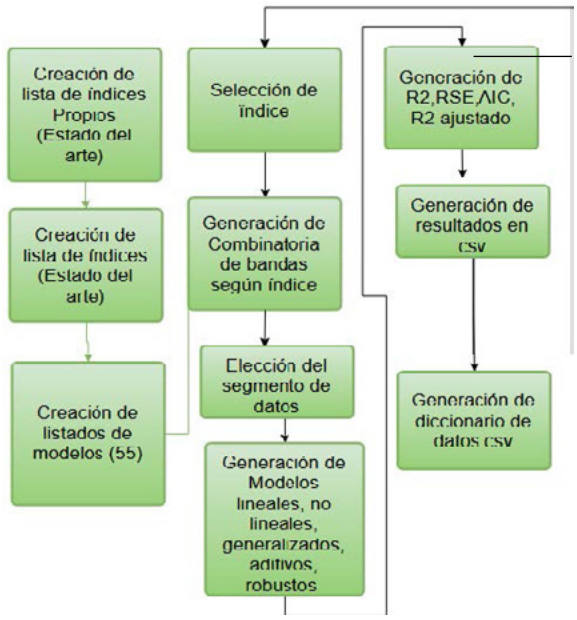

Fig. 4. Algoritmo de generación de resultados: R^2, RSE, AIC, R^2 ajustado.

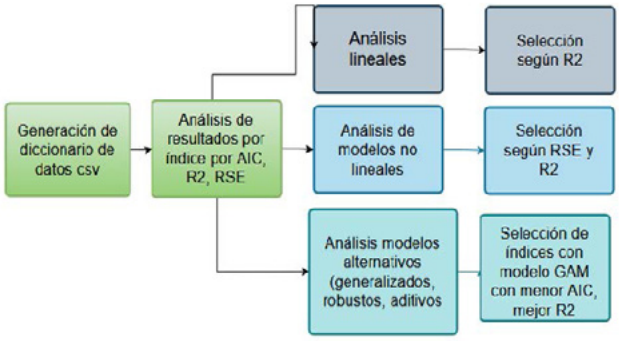

Fig. 5. Análisis de modelos lineales, no lineales, alternativos.

Se analizó el listado comparativo de índices (figura 5) generados sistemáticamente para elegir los candidatos, ubicando los modelos que tenían menor AIC, y mayor R^2, dentro de esta sección de modelos especiales, se encuentran modelos robustos, y funciones no lineales, se analizaron los modelos aditivos generalizados (GAM), se observó en ellos mejores resultados en AIC y R^2 que los otros modelos probados.

Para el modelo GAM ganador del primer listado se generó un lazo algorítmico para un total de 10 posibles modelos candidatos, con parámetros diferentes de suavizamiento, nudos y correlación, se evaluaron los resultados mediante la evaluación de los valores de AIC y R^2.

Finalmente se estimó el índice TBDO contra el índice candidato, con los mismos parámetros del modelo GAM.

Para la prueba de Análisis espacial (figura. 6) se utilizó: la evaluación de la gráfica de residuos versus valores estimados de la regresión para detectar patrones de dependencia como principal referencia para encontrar el incumplimiento de independencia.

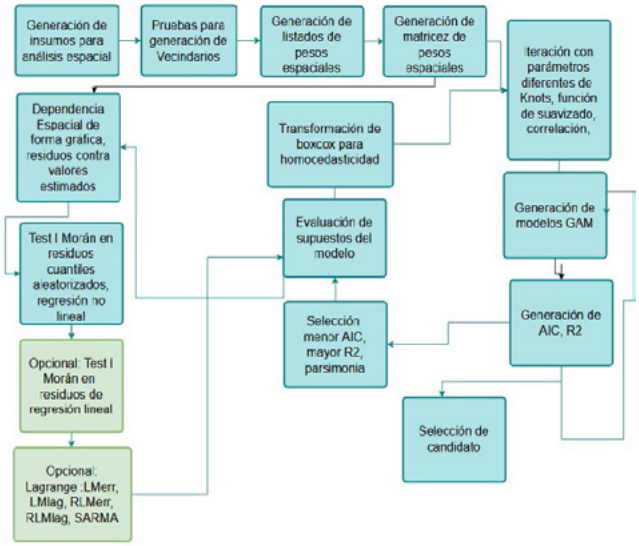

Fig. 6. Análisis espacial.

La utilización de método de residuos cuantiles aleatorios en la prueba de I Moran mediante el paquete DHARMa de R. Se comparó el resultado con la salida de la prueba de residuales de I Moran, en el caso de la regresión cúbica, esta prueba se utiliza para relaciones lineales, el estado del arte señala [5] que para casos no lineales dicha prueba falla, en este caso pese a que la gráfica demostraba un patrón, la prueba no arrojó evidencias para rechazar la hipótesis nula de independencia incluyendo Multiplicadores de Lagrange: LMerr, LMlag, RLMerr, RLMlag, SARMA. Se agregó latitud y longitud al modelo GAM para obtener mejor explicación de desvianza.

III. RESULTADOS

Los mejores modelos encontrados en este estudio se muestran en la tabla 2, la figura 7 muestra las bandas utilizadas en el modelo GAM, la figura 8

muestra la gráfica ajustada del índice cúbico y la clorofila-a.

Tabla 2.: Resumen de los modelos resultantes

Modelo	R^2	AIC	Supuestos superados
Empirico: $Chl - a = B5 * B3 * B10$	0.8145	127.26	Independencia Homocedasticidad Normalidad
GAM: $(Chl - a)^{1/2} = s(\frac{B8 - B6}{B1},$ *fx = FALSE, k = 15, bs = "cr"+ s(lon, lat, bs = "ts", k = 7)*	0.986	48.53	Independencia Gamma Linealidad
Cúbico: $Chl - a = \frac{B8 * B12}{B8A}$	0.854	423.71	Normalidad Homocedasticidad Linealidad
TBDO: $In\,(chl - a) = B6 * (\frac{1}{B4} - \frac{1}{B5})$	0.587	97.05	Independencia Homocedasticidad Linealidad

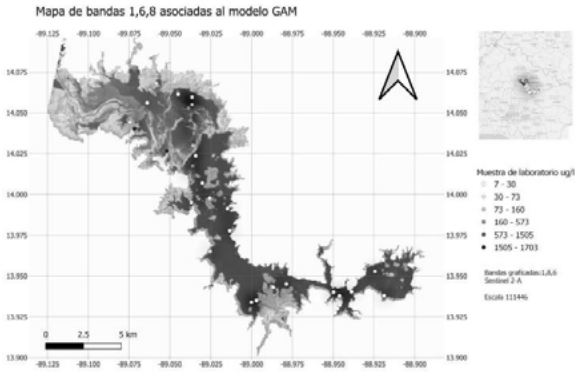

Fig. 7. Mapa de las bandas 1, 6 y 8 asociadas al modelo GAM.

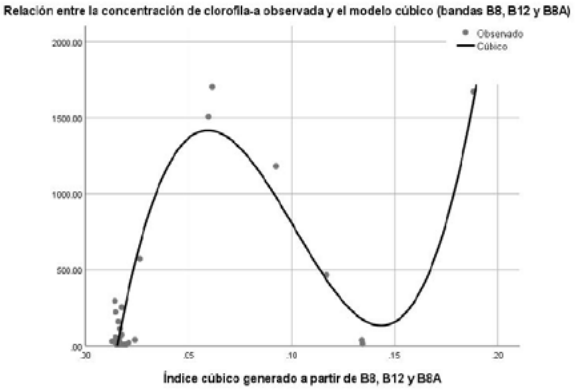

Fig. 8. Gráfica ajustada del índice cúbico y la clorofila-a.

IV. DISCUSIÓN

En esta investigación se encontraron tres modelos que al compararlos contra el TBDO ofrecen mejor estimación de la concentración de clorofila-a en el Humedal. Las relaciones encontradas fueron complejas del tipo no lineal y los modelos hicieron uso de bandas espectrales no utilizadas en el estado del arte [6], incluso con resoluciones espaciales más grandes como la banda 1. Los R^2 encontrados están arriba de 0.80, siendo este un buen resultado con respecto a modelos encontrados en otros cuerpos de agua. En esta investigación también se utilizó la evaluación del menor AIC como criterio de selección, ya que evalúa la eficacia de un modelo cuantificando la pérdida de información que este produce al representar la realidad subyacente; por lo tanto, un modelo es mejor cuanto menor sea esta pérdida [7].

El impacto de la no linealidad obligó a hacer uso de técnicas de análisis espacial de alta complejidad, por lo que no se pudo comparar la correlación espacial de los residuos de las regresiones a través de los métodos de Lagrange o métodos similares.

Las teorías encontradas para medir la concentración de clorofila-a por medio de las bandas satelitales presentan en su mayoría un modelo de regresión lineal basado en el R^2 y AIC. El índice TBDO ha dado buenos resultados en estudios anteriores [6] alcanzando por ejemplo un R^2 de 0.6377 y una pendiente cercana a uno al ser aplicado con imágenes Sentinel-2 en la Albufera de Valencia [8] y un R^2 de 0.89 en el estudio de validación del algoritmo siempre en la zona de la Albufera de Valencia [8].

El objetivo de esta investigación fue comparar distintos índices con todas las posibles combinaciones de bandas, sin tomar en cuenta si se utilizaban o no las bandas recomendadas por la literatura [6], para medir la concentración de clorofila-a. Al comparar más de 100,000 modelos, el que obtuvo R^2 más alto y AIC más bajo de los modelos polinomiales, fue un modelo cúbico con las bandas B8, B12 y B8A.

El mejor índice cúbico tuvo un R^2 de 0.854, con ANOVA y coeficientes significativos. Es importante señalar, que mientras la banda B8 es comúnmente utilizada en la literatura para estimar la clorofila-a, la inclusión de la banda 8A y la banda B12 no es habitual en estudios similares. Su presencia en este modelo sugiere la posibilidad de explorar combinaciones de bandas alternativas a la tradicionalmente empleadas, por ejemplo, la combinación de las bandas 5 y 8A han demostrado en trabajos recientes [9].

El TBDO con un R^2 de 0.3239 de una regresión lineal y las bandas B4, B5 y B6, tuvo coeficientes significativos. El R^2 del TBDO predice poco la concentración de clorofila-a del humedal, al hacer una

transformación logarítmica al TBDO se obtuvo un R^2 de 0.58, por lo cual el modelo empírico utilizado en el estudio logró estimar mejor la concentración de clorofila-a del humedal.

Al evaluar el índice TBDO en un modelo GAM se obtuvo una medida equivalente al R^2 de 0.66 y un AIC de 116.3157 siendo este menor que el GAM propuesto (tabla 2).

Para evaluar la dependencia espacial en los residuos de los modelos, se utilizó la búsqueda de patrones en la salida de residuales, con el gráfico: residuales versus valores estimados. Siendo el índice de Moran calculado con los residuos cuantiles aleatorios, prueba de regresión de prueba de Moran y multiplicadores de Lagrange, obteniendo los mejores resultados con visualización de los residuos y el índice de I Moran calculado con los residuos cuantiles.

Los resultados obtenidos con el índice cúbico con las bandas B8, B12 y B8A, tuvieron un R^2 bastante aceptable para medir la concentración de clorofila-a del humedal, pero el grado de la regresión fue alto y al aumentar el orden del polinomio, la matriz con la cual se calculan los estimadores de los parámetros se vuelve mal acondicionada y esto produce la posibilidad de aumentar el error de predicción, hay que tener en consideración lo anterior a la hora de elegir este modelo para realizar predicciones.

La relación entre las bandas del satélite y el valor de la concentración de la clorofila-a en el humedal no siguen un patrón lineal.

Aquellas bandas que no estaban contempladas en el estado del arte [6] como la B1, han sido utilizadas en esta investigación pese a la pérdida de la calidad de la resolución de 20 metros cuadrados por la de 60 metros cuadrados y no tener relación con las bandas infrarrojas asociadas a la vegetación.

Utilizar modelos GAM con la incorporación de la latitud y longitud espacial permitieron junto a la banda B1 estimar los valores abajo de 400 ug/l en la zona alta norte y zona sur.

La necesidad de más investigación en el área de estadística espacial específicamente para evaluar los supuestos de dependencia de los residuos para relaciones no lineales, es de gran importancia para futuras investigaciones, ya que se tuvo que hacer uso de métodos no tradicionales y utilizar paquetes creados en 2021 en software R (DHARMa) para distribuciones no normales, con variables categóricas dependientes, para poder estimar el valor de I Moran de los residuos, con la técnica de residuos cuantiles aleatorios.

En cuanto al uso de modelos de error y retraso espacial, sugeridos por el estado del arte [10] para los nuevos estudios de modelado de la calidad del agua, no se logró utilizar dichos métodos debido a la relación no lineal entre los índices y el valor de la concentración de la clorofila-a, por lo que los resultados de la investigación se mantienen en congruencia con el estado del arte en el aspecto de no utilizar dichos modelos espaciales para estos casos. A excepción del modelo cúbico los otros índices superaron el supuesto de independencia de los residuos.

Debido a las bandas encontradas y con base a los estudios realizados previamente en el humedal, queda abierto el tema para los investigadores de las ciencias ambientales y teledetección de encontrar las causas de dicho comportamiento revelado por los modelos.

V. CONCLUSIONES

Se comprobó con tres modelos de regresión que el uso de la teledetección es una opción para estimar la concentración de clorofila-a en el humedal y puede usarse para que los especialistas en ciencias ambientales utilicen dichos valores estimados en sistemas de alerta temprana asociados a la contaminación en el cuerpo de agua.

Se verificó que la concentración de clorofila-a en el humedal del Cerrón Grande pudo ser estimada por medio de modelos de regresión, obteniendo mejores resultados que el TBDO.

Con respecto a la evaluación de la independencia de los residuos, el modelo cúbico fue el único que presentó autocorrelación espacial, la cual no pudo ser corregida debido a la complejidad no lineal del modelo y para detectarla se tuvo que usar una versión avanzada del I Moran, para la cual se sugiere mayor investigación.

El análisis espacial ha permitido una mayor comprensión del fenómeno de la clorofila-a y su relación con las bandas del satélite Sentinel-2A a lo largo y ancho del humedal.

Pese a que la literatura encontrada sugiere usar las bandas dentro y cercanas al infrarrojo, los resultados de esta investigación ponen en discusión la utilización de otras bandas diferentes.

Con respecto a la resolución espacial proporcionada por el satélite Sentinel-2A, la mayoría de las modelos resultantes de esta investigación utilizaron bandas con una resolución menor o igual a 20 metros cuadrados, logrando sacar provecho de la alta resolución que provee este satélite.

El haber utilizado la banda 1, que discrimina la costa y aerosol, incrementó la resolución espacial mínima de 20 a 60 metros cuadrados, dio excelentes resultados.

Los modelos propuestos en esta investigación, pese a que por la complejidad del fenómeno de estudio en su mayoría no cumplen con todos los supuestos de validación estadística, pueden ser utilizados en la práctica.

El análisis de componentes principales fue una técnica adecuada para comprobar la relación lineal entre las bandas del infrarrojo y cercanas a ella con la clorofila-a.

VI. AGRADECIMIENTOS

A la directora de la Maestría en Estadística Aplicada a la Investigación Dra. Lorena Ivon Rivas de Mendoza por su apoyo y coordinación.

A las investigadoras de la UCA Dra. María Dolores Rovira, Inga. Ingrid García, Metzi Aguilar e Inga. Astrid González por su colaboración con los datos y seguimiento en el proceso.

A la comunidad de software libre y open source involucrada. en el desarrollo de paquetes de R y QGIS.

A los docentes de la Maestría.

A la asistente administrativa Paola G. Pimentel.

Y a toda la comunidad UCA.

VII. REFERENCIAS

[1] R. Duarte, Relación entre clorofila superficial y clorofila en la zona eufótica del golfo de California: Posible aplicación para estimar la producción primaria con datos obtenidos por sensores remotos. Ciencias Marinas, 19(4), 473-490 1993.

[2] D. Russin, P. ten Brink, A. Farmer, T. Badura, D. Coates, J. Förster y N. Davidson, The economics of ecosystems and biodiversity for water and wetlands. IEEP, 78, 2012.

[3] M. Rovira, J. Ortez, L. Morán, G. Arevalo, J. Franco y D. Linares, Establecimiento de línea base para la identificación de Cianobacterias potencialmente tóxicas del Embalse Cerrón Grande, 2020.

[4] S. Monge, Desarrollo del método para la cuantificación de la Clorofila-a en muestras de agua, por espectroscopia ultravioleta visible, 2015.

[5] F. López-Hernández, A. Artal-Tur y M. Maté-Sánchez-Val, Identifying nonlinear spatial dependence patterns by using non-parametric tests. Investigaciones Regionales (21), 19-36. (41), 37-47, 2011.

[6] A. Gitelson, G. Dall'Olmo, W. Moses, D. Rundquist, T. Barrow, T. Fisher, y J. Holz, Un modelo semianalítico simple para la estimación remota de clorofila-a en aguas turbias: Validación. Detección remota del medio ambiente, 112 (9), 3582-3593, 2008.

[7] FasterCapital, Criterio de información de Akaike (AIC): el criterio secreto para la selección de modelos, 2025, [Online] Available: https://fastercapital.com/es/contenido/Criterio-de-informacion-de-Akaike-AIC-AIC-El-criterio-secreto-para-la-seleccion-de-modelos.html#-Qu-es-el-criterio-de-informaci-n-de-Akaike-AIC-

[8] P. Urrego, X. Sòria-Perpinyà, J. Delegido, M. Pereira-Sandoval, A. Ruiz-Verdu, C. Tenjo, E. Vicente, Eduardo, J. Soria y J. Moreno, Monitoreo ecológico multitemporal de la Albufera de Valencia con imágenes Sentinel-2, 2018, [Online] Available: https://www.researchgate.net/publication/328783570_Monitoreo_ecologico_multitemporal_de_la_Albufera_de_Valencia_con_imagenes_Sentinel-2

[9] W. Jang, J. Kim, J.H. Kim, J. K. Shin, K. Chon, E.T. Kang, Y. Park, y S. Kim, Evaluation of Sentinel-2 based chlorophyll-a estimation in a small-scale reservoir: Assessing accuracy and availability. Remote Sensing, 16(2), 315, 2025, https://doi.org/10.3390/rs16020315

[10] L. Miralha, y D. Kim, D, Accounting for and predicting the influence of spatial autocorrelation in water quality modeling. ISPRS international Journal of Geoinformation, 64-64, 2018.