UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

PROYECTO MAESTRÍA

Estimación de Parámetros para la Evaluación del Índice de Calidad de Agua en el Lago George Mediante el Uso de Imágenes Sentinel-2 y Machine Learning.

Autor:
Gerardo Andrés Patiño Montoya.

Director:
PhD. Rubén Javier Medina Daza.



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Grupo de Investigación:
Grupo de Estudio en Temas de la Física, de la Estadística y de la Matemática (GEFEM).

Modalidad: Profundización.

Línea de Investigación:
Procesamiento Digital de Imágenes y Fotogrametría.

Maestría en Ciencias de la Información y las Comunicaciones.

Énfasis: Geomática.

31 de julio de 2026

Índice general

Introducción **1**

Resumen **3**

Abstract **4**

1. Objetivos **5**
1.1. Objetivo general . . . 5
1.2. Objetivos específicos . . . 5

2. Justificación. **6**

3. El Problema de Investigación **8**
3.1. Planteamiento del Problema. . . . 9
3.2. Formulación del Problema. . . . 9
3.3. Sistematización del Problema. . . . 10

4. Estado del Arte. **11**

5. Marco Teórico. **14**
5.1. Parámetros de Calidad del Agua. . . . 14
5.1.1. Fósforo Total. . . . 14
5.1.2. Nitrógeno Total. . . . 15
5.1.3. Clorofila-a. . . . 15
5.1.4. Profundidad del Disco Secchi. . . . 15
5.1.5. Solidos Suspendidos Totales. . . . 15
5.1.6. Cloruro. . . . 15
5.1.7. Oxígeno Disuelto. . . . 16
5.1.8. Nivel de Acidez (pH). . . . 16
5.1.9. Temperatura superficial. . . . 16
5.1.10. Turbidez. . . . 16
5.1.11. Arsénico. . . . 17
5.1.12. Dureza. . . . 17
5.1.13. Potasio. . . . 17
5.1.14. Sodio. . . . 17
5.1.15. Sulfato. . . . 18
5.2. Índice de Calidad del Agua. . . . 19
5.2.1. Reseña Histórica de los Índices de Calidad de Agua Predecesores al CCME WQI. 19
5.2.2. Centro de St Laurent. . . . 19
5.2.3. Quebec. . . . 20

5.2.4. British Columbia. . . . 20
5.2.5. Alberta. . . . 20
5.2.6. Ontario. . . . 20
5.2.7. Modelo Conceptual del CCME WQI. . . . 21
5.2.8. Formulación Final del CCME WQI. . . . 21
5.2.9. Categorización del CCME WQI. . . . 23
5.3. Programa Sentinel-2. . . . 24
5.3.1. Características Generales. . . . 24
5.3.2. Características del Sensor. . . . 24
5.3.3. Tipos de Productos. . . . 25
5.3.4. Nomenclatura de los Productos. . . . 26
5.4. Regresión Lineal Múltiple por MCO. . . . 27
5.4.1. Coeficiente de Determinación Múltiple. . . . 27
5.4.2. Supuestos de Normalidad. . . . 28
5.4.3. Supuesto de Homocedasticidad. . . . 28
5.4.4. Supuesto de No Autocorrelación. . . . 29
5.4.5. Supuesto de Multicolinealidad. . . . 29
5.4.6. Prueba de Significancia General de una Regresión Múltiple. . . . 30
5.5. Regresión Lineal Múltiple por GLSAR. . . . 31
5.5.1. Estructura del modelo. . . . 31
5.5.2. Matriz de Covarianza HAC. . . . 31
5.6. Modelo MARS. . . . 33
5.6.1. Comparación con las Regresiones Lineales y No Lineales Convencionales. . . . 33
5.6.2. Estructura Básica del Modelo. . . . 33
5.6.3. Método de Selección de Funciones. . . . 35

6. Diseño Metodológico. 37
6.1. Zona de Estudio. . . . 37
6.1.1. Características Físicas. . . . 37
6.1.2. Topografía y Vida Acuática. . . . 37
Uso e Importancia Económica. . . . 38
6.2. Fase 1: Preprocesamiento de los datos y obtención de las ecuaciones de los parámetros. 40
6.3. Fase 2: Estimación de los parámetros y modelado del índice de calidad del agua. . . . 41
6.4. Fase 3: Evaluación del modelo. . . . 41
6.5. Fase 4: Validación del modelo. . . . 41

7. Descripción y Análisis Exploratorio de Datos. 42
7.1. Parámetros de Calidad del Agua. . . . 42
7.1.1. Estadísticas descriptivas. . . . 45
7.1.2. Histogramas. . . . 46
7.1.3. Diagrama *Box-Plot*. . . . 48
7.2. Imágenes Sentinel-2. . . . 51

8. Desarrollo de los Modelos de Regresión Lineal. 53
8.1. Limpieza y Filtrado de Variables Significativas. . . . 53
8.1.1. Limpieza de los Datos. . . . 54
8.1.2. Selección de Variables Significativas. . . . 55
8.1.3. Ajuste de los Modelos Preliminares. . . . 56

8.1.4. Análisis de Observaciones Influyentes. . . . 56
8.2. Ajuste de los Modelos Finales. . . . 58
8.2.1. Arsénico. . . . 58
8.2.2. Sulfato. . . . 60
8.2.3. Clorofila-a. . . . 62
8.2.4. Cloruro. . . . 64
8.2.5. Oxígeno disuelto. . . . 66
8.2.6. Dureza. . . . 68
8.2.7. Fósforo. . . . 70
8.2.8. Nitrógeno. . . . 72
8.2.9. Acidez (pH). . . . 73
8.2.10. Potasio. . . . 75
8.2.11. Profundidad del disco Secchi. . . . 77
8.2.12. Sodio. . . . 79
8.2.13. Temperatura. . . . 81
8.2.14. Sólidos suspendidos totales. . . . 83
8.2.15. Turbidez. . . . 84
8.3. Resumen de los modelos finales. . . . 86

9. Predicción de Parámetros. 88
9.1. Validación de Rangos y Extracción de Valores de Reflectancia. . . . 88
9.2. Predicciones. . . . 93

10. Cálculo del Índice de Calidad del Agua. 95
10.1. Fase 1: Cálculo de los Factores por Parámetro. . . . 95
10.1.1. Arsénico. . . . 95
10.1.2. Sulfato. . . . 96
10.1.3. Clorofila-a. . . . 97
10.1.4. Cloruro. . . . 97
10.1.5. Dureza. . . . 98
10.1.6. Acidez (pH). . . . 99
10.1.7. Sodio. . . . 99
10.1.8. Temperatura. . . . 100
10.1.9. Sólidos suspendidos totales. . . . 101
10.1.10. Turbidez. . . . 101
10.2. Fase 2: Agrupación de los Factores y Cálculo del Índice. . . . 102
10.2.1. Distribución del WQI. . . . 103
10.3. WQI en las Estaciones de Referencia. . . . 106

11. Implementación del Modelo MARS. 108
11.1. División de Épocas Estacionales. . . . 108
11.1.1. Invierno. . . . 108
11.1.2. Primavera. . . . 109
11.1.3. Verano. . . . 110
11.1.4. Otoño. . . . 110
11.2. Mejor Modelo. . . . 111
11.2.1. Primera iteración. . . . 111
11.2.2. Segunda iteración. . . . 112

11.2.3. Dominio espectral. 112

12. Evaluación del Modelo MARS. 114
12.1. Escenario 1: Julio. 114
12.2. Escenario 2: Agosto. 118
12.3. Escenario 3: Septiembre. 120
12.4. Resumen de los Escenarios. 122

13. Validación del Modelo MARS. 123
13.1. Métricas de Concordancia. 123
13.1.1. Índice de Concordancia (*Index of Agreement*). 123
13.1.2. Índice de Dispersión (*Scatter Index*). 124
13.2. Estadísticas de los Residuos. 124
13.2.1. Estadísticas descriptivas. 124
13.2.2. Estadísticas de validación. 125
13.3. Distribución de los Residuos. 126
13.3.1. Distribución espacial. 127
13.4. Prueba de Autocorrelación Espacial. 130
13.5. Bootstrap y Validación Cruzada. 130
13.6. Prueba de Significancia. 132

14. Conclusiones. 134

15. Principales Aportes. 137

16. Recomendaciones. 139

Bibliografía 140

Índice de figuras

5.1. Modelo conceptual del CCME WQI. . . . 21
5.2. Características técnicas del sensor. . . . 24
5.3. Cobertura espectral del Sentinel-2. . . . 26
5.4. Prueba de hipótesis de normalidad. . . . 28
5.5. Comportamiento de homocedasticidad. . . . 29
5.6. Ejemplo de residuos correlacionados. . . . 30
5.7. Ejemplo de residuos no autocorrelacionados. . . . 30
5.8. Ejemplos del comportamiento de la línea de regresión en datos con distribución polinómica de grado 3. . . . 34
5.9. Ejemplos de conjuntos de datos con varios nudos. . . . 35
5.10. Arquitectura del modelo MARS. . . . 36

6.1. Localización del lago George. . . . 38
6.2. Metodología propuesta para la evaluación del índice de calidad del agua mediante el uso de imágenes Sentinel-2 y *machine learning*. . . . 39

7.1. Ubicación de las estaciones en el Lago George 44

9.1. Ubicación de las estaciones en el Lago George 91
9.2. Distribución de los evaluables sobre el lago George. Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023). . . . 92

10.1. Distribución del WQI en los puntos evaluables. . . . 104
10.2. Clasificación del WQI sobre el lago George. Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023). . . . 105

12.1. Área efectiva para la evaluación del modelo MARS en la zona occidental del lago George. Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023). . . . 115
12.2. Estimación del WQI para el mes de julio en la zona occidental del lago George. Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023). . . . 117
12.3. Estimación del WQI para el mes de agosto en la zona occidental del lago George. Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023). . . . 119
12.4. Estimación del WQI para el mes de septiembre en la zona occidental del lago George. Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023). . . . 121

13.1. Histograma de los residuos del modelo MARS. . . . 126
13.2. Distribución de los residuos del modelo MARS. . . . 127
13.3. Distribución de los residuos del modelo MARS en el lago George. Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023). . . . 129

Índice de tablas

5.1. Categorías del índice de calidad del agua según el CCME WQI. 23
5.2. Bandas del programa Sentinel-2: Longitudes de onda y resolución espacial. 25
5.3. Nomenclatura del producto del programa Sentinel-2. 26

7.1. Estadísticas descriptivas de los parámetros medidos en el Lago George. 45
7.2. Histogramas de los parámetros medidos en el Lago George. 46
7.3. *Box-Plot* de los parámetros medidos en el Lago George. 49
7.4. Producto y fecha de adquisición de las imágenes. 51

8.1. Datos atípicos eliminados por parámetros. 54
8.2. Datos atípicos e influyentes eliminados en cada parámetros. 57
8.3. Resumen del modelo de regresión lineal por MCO para el arsénico. 58
8.4. Resumen del modelo de regresión lineal por GLSAR para el arsénico. 59
8.5. Resumen del modelo de regresión lineal por MCO para el sulfato. 60
8.6. Resumen del modelo de regresión lineal por GLSAR para el sulfato. 61
8.7. Resumen del modelo de regresión lineal por MCO para la clorofila. 62
8.8. Resumen del modelo de regresión lineal por GLSAR para la clorofila. 63
8.9. Resumen del modelo de regresión lineal por MCO para el cloruro. 64
8.10. Resumen del modelo de regresión lineal por GLSAR para el cloruro. 65
8.11. Resumen del modelo de regresión lineal por MCO para el oxígeno disuelto. 66
8.12. Resumen del modelo de regresión lineal por GLSAR para el oxígeno disuelto. 67
8.13. Resumen del modelo de regresión lineal por MCO para la dureza. 68
8.14. Resumen del modelo de regresión lineal por GLSAR para la dureza. 69
8.15. Resumen del modelo de regresión lineal por MCO para el fósforo. 70
8.16. Resumen del modelo de regresión lineal por GLSAR para el fósforo. 71
8.17. Resumen del modelo de regresión lineal por MCO para el nitrógeno. 72
8.18. Resumen del modelo de regresión lineal por MCO para el pH. 73
8.19. Resumen del modelo de regresión lineal por GLSAR para el pH. 74
8.20. Resumen del modelo de regresión lineal por MCO para el potasio. 75
8.21. Resumen del modelo de regresión lineal por GLSAR para el potasio. 76
8.22. Resumen del modelo de regresión lineal por MCO para la SDD. 77
8.23. Resumen del modelo de regresión lineal por GLSAR para la SDD. 78
8.24. Resumen del modelo de regresión lineal por MCO para el sodio. 79
8.25. Resumen del modelo de regresión lineal por GLSAR para el sodio. 80
8.26. Resumen del modelo de regresión lineal por MCO para la temperatura. 81
8.27. Resumen del modelo de regresión lineal por GLSAR para la temperatura. 82
8.28. Resumen del modelo de regresión lineal por MCO para los sólidos suspendidos totales. 83
8.29. Resumen del modelo de regresión lineal por MCO para la turbidez. 84
8.30. Resumen del modelo de regresión lineal por GLSAR para la turbidez. 85

8.31. Resumen de los modelos significativos. 86
8.32. Resumen de los modelos no significativos. 87

9.1. Rango de las variables de cada modelo. 88
9.2. Puntos válidos por parámetro. 89
9.3. Cantidad de muestras finales. 93
9.4. Cantidad de muestras finales. 94

10.1. Resumen de las pruebas fallidas del arsénico. 95
10.2. Resumen de las pruebas fallidas del sulfato. 96
10.3. Resumen de las pruebas fallidas de la clorofila-a. 97
10.4. Resumen de las pruebas fallidas del cloruro. 98
10.5. Resumen de las pruebas fallidas del cloruro. 98
10.6. Resumen de las pruebas fallidas del pH. 99
10.7. Resumen de las pruebas fallidas del sodio. 100
10.8. Resumen de las pruebas fallidas de la temperatura. 100
10.9. Resumen de las pruebas fallidas de los TSS. 101
10.10Resumen de las pruebas fallidas de la turbidez. 102
10.11Resumen de las excursiones para el factor de la amplitud. 102
10.12Resumen de las excursiones para el factor de la amplitud. 103
10.13Resumen de las excursiones del WQI en las estaciones de referencia. 106
10.14WQI calculado para las estaciones. 107

11.1. Desempeño del modelo MARS en diciembre. 109
11.2. Desempeño del modelo MARS en marzo. 110
11.3. Desempeño del modelo MARS en julio. 110
11.4. Desempeño del modelo MARS en octubre. 110
11.5. Desempeño del modelo MARS en julio solo con las bandas más significativas. 112
11.6. Estadísticas descriptivas de las bandas del MARS para el mes de julio. 113

12.1. Resumen de la evaluación del modelo MARS en la zona occidental del lago George para el mes de julio. 116
12.2. Resumen de la evaluación del modelo MARS en la zona occidental del lago George para el mes de agosto. 118
12.3. Resumen de la evaluación del modelo MARS en la zona occidental del lago George para el mes de septiembre. 120
12.4. Estadísticas descriptivas de la evaluación del modelo MARS en los tres escenarios propuestos. 122

13.1. Resumen de las estadísticas descriptivas de los residuos. 125
13.2. Resumen de las estadísticas de validación de los residuos. 126
13.3. Resultados de la prueba I de Moran para la detección de autocorrelación espacial. 130
13.4. Resultados de la *Bootstrap* para el conjunto de datos completo. 131
13.5. Resultados de la validación cruzada. 132
13.6. Resultados de la prueba F para la significancia estadística del modelo. 133

Introducción

La gestión sostenible de los recursos hídricos y de la preservación de los ecosistemas acuáticos requiere un plan de monitoreo constante y eficiente de la calidad del agua en los diferentes cuerpos que constituyen una determinada región. Desde un punto de vista físico-químico, este proceso, tradicionalmente, se ha llevado a cabo por medio de análisis de muestreo en campo, es decir, teniendo contacto directo con la superficie en evaluación. Si bien esta técnica ofrece resultados precisos y confiables, también representa un costo elevado y una compleja planificación debido a los diferentes obstáculos que se deben enfrentar al momento de visitar un lugar, ya sea por su difícil acceso, por las circunstancias ecológicas (por ejemplo, un cuerpo de agua que sirva de hábitat de especies animales consideradas salvajes) o la limitación jurídica de espacios restringidos por ser privados. En consecuencia, el uso de tecnologías de análisis geográfico como el uso de imágenes satelitales ofrece una alternativa comparativamente eficaz y amigable en términos de costos, enfocada en la complementación y optimización en el monitoreo de la calidad del agua.

Por otra parte, una de las limitaciones más grandes con relación al uso de imágenes satelitales viene dada por la resolución del sensor que se elija para el desarrollo del análisis; por tal motivo, los productos provenientes de un programa como el Sentinel-2 representan un recurso valioso para la observación remota de la superficie terrestre debido a que éste cuenta con una resolución espacial de 10 metros, aportando así una mayor captura de información con respecto a sensores como Landsat-8, sensor famoso por ser utilizado como insumo para diferentes tipos de investigaciones relacionadas con la teledetección. Asimismo, para la fecha de realización de la presente investigación, las imágenes de Sentinel pueden ser descargadas de manera gratuita directamente de la página de Copérnicus, permitiendo así el procesamiento y obtención de parámetros físico-químicos del agua a través del análisis de la información espectral capturada por el satélite sin representar un gasto mayor. Adicionalmente, la integración de modelos de *machine learning* con la teledetección abre nuevas oportunidades para el modelado y predicción del comportamiento de múltiples variables que inciden directa e indirectamente en la calidad del agua, mejorando su evaluación y monitoreo en términos de costos, tiempos y frecuencias.

En cuanto a la zona de estudio, el estado de Florida, desde el inicio del presente siglo, ha estado viéndose amenazado por el desabastecimiento del recurso hídrico debido a diferentes factores relacionados con el uso indiscriminado, la sobrepoblación y la contaminación irreversible del agua. Ante este panorama, la búsqueda de fuentes alternativas se ha hecho una de las misiones principales de investigadores y académicos. No obstante, para que la extracción del recurso sea efectiva, es necesario determinar si el agua es apta o no para ser distribuida y consumida por la población; es así que el monitoreo y posterior análisis de la calidad del agua cobran tal relevancia. En este contexto, se ha elegido el lago George como caso de estudio, ya que este cuerpo no ha sido objeto de evaluación por medios remotos, lo que lo convierte en una localización idónea.

En síntesis, el modelo propuesto permitirá evaluar la calidad del agua del lago George mediante el índice promulgado por el Consejo Canadiense de Ministros de Medio Ambiente (CCME) el cual se basa en los parámetros físico-químicos que van a ser estimados bajo la aplicación de regresiones

multivariadas, analizando la correspondencia espectral de las variables con respecto a cada una de las bandas que componen las imágenes satelitales del programa Sentinel-2. La aplicación del modelo, de igual forma, posibilitará validar la efectividad de la integración de modelos de *machine learning* con percepción remota, contribuyendo así al fortalecimiento de la innovación de metodologías para la gestión y preservación de recursos hídricos, así como a la búsqueda de fuentes alternativas de agua en sitios en ambientes apremiantes ante la falta o disminución del líquido preciado.

Resumen

Con el fin de implementar una metodología alternativa para el monitoreo de la calidad del agua en el lago George, se desarrollaron modelos de regresión lineal para estimar parámetros de calidad hídrica a partir de mediciones de septiembre de 2022 a abril de 2024 como variables dependientes, y una combinación de bandas, índices y cocientes espectrales del programa Sentinel-2 como variables independientes. De los conjuntos analizados, 10 lograron un ajuste aceptable, con un R^2 ajustado mayor al 70 %; no obstante, debido a la marcada autocorrelación temporal en los residuos, fue necesario realizar el ajuste por mínimos cuadrados generalizados con autocorrelación (GLSAR), lo que permitió obtener modelos estadísticamente válidos.

A partir de las estimaciones de los modelos, se aplicó el índice de calidad del agua del Consejo Canadiense de Ministros del Medio Ambiente. Posteriormente, se ajustó un modelo de regresión de splines adaptativos multivariada (MARS) entre el índice obtenido y las bandas del programa Sentinel-2, divididas en cuatro imágenes representativas de cada estación del año. El análisis reveló que la imagen de julio presentó un mejor desempeño, con un RMSE de 2,76 y un error absoluto de 1,99. Aun así, el mejor modelo solo pudo explicar 0,9 % de la variabilidad del índice en el lago, debido a que los valores estaban predominantemente concentrados entre 80 y 85; por lo que no obtuvo una variabilidad que pudiera ser explicada por el MARS, resultando en un ajuste poco satisfactorio.

Abstract

To implement an alternative methodology for water quality monitoring in Lake George, linear regression models were developed to estimate water quality parameters using field measurements collected between September 2022 and April 2024 as dependent variables and a combination of Sentinel-2 spectral bands, indices, and band ratios as predictors. Of the datasets analyzed, ten achieved satisfactory performance, with an adjusted R^2 greater than 70 %. However, due to the strong temporal autocorrelation detected in the residuals, generalized least squares with autoregressive errors (GLSAR) was employed instead of ordinary least squares, resulting in statistically valid models.

The estimated parameters were subsequently used to calculate the Canadian Council of Ministers of the Environment Water Quality Index. A multivariate adaptive regression splines (MARS) model was then fitted to relate the calculated index to Sentinel-2 spectral bands from four representative images corresponding to the different seasons of the year. The July image provided the best predictive performance, with an RMSE of 2.76 and a mean absolute error (MAE) of 1.99. Nevertheless, the best model explained only 0.9 % of the variability in the water quality index because most WQI values were concentrated between 80 and 85, indicating limited variability in the response variable and, consequently, poor predictive performance of the MARS model.

1

Objetivos

1.1. Objetivo general

Desarrollar un modelo que evalúe el índice de calidad del agua en el lago George mediante la estimación de parámetros usando imágenes Sentinel-2 y una regresión de splines adaptativos multivariada para el año 2023.

1.2. Objetivos específicos

- Implementar un algoritmo de regresión multivariada para la estimación de los parámetros iniciales mediante la combinación de bandas de imágenes satelitales.
- Generar un modelo a partir de una regresión de splines adaptativos multivariada capaz de estimar el índice de calidad del agua propuesto por el Consejo Canadiense de Ministros del Medio Ambiente con base en el comportamiento espectral de los parámetros estimados.
- Evaluar el índice de calidad del agua mediante la aplicación del modelo propuesto en el Lago George.
- Validar el modelo mediante criterios de verificación estadística para corroborar el desempeño del índice de calidad de agua estimado en la zona de estudio.

2

Justificación.

El análisis de la calidad del agua ha tenido avances considerables en las últimas décadas gracias al desarrollo de diferentes técnicas para la estimación de parámetros e índices correspondientes. No obstante, algunas de estas técnicas se han hecho basándose en modelos o ecuaciones ya propuestas (Ahmed et al., 2022b), (Ahmed et al., 2022a), las cuales se enfocan en la aplicación de modelos y técnicas de *machine learning* sobre datos generados a partir de ecuaciones previamente ajustadas por otros autores en zonas con características similares. Si bien esta técnica se ha popularizado a lo largo de los años, no se puede soslayar el hecho de que el sesgo de los resultados va a estar basado en la temporalidad y precisión de los datos utilizados para generar dichas ecuaciones, lo que puede ocasionar una estimación errada al tratarse de cuerpos de agua distintos. En ese orden de ideas, la presente investigación pretendió desarrollar un enfoque práctico orientado a la generación de ecuaciones que permitieran estimar los parámetros de calidad del agua con base en información espectral tomada directamente de la zona de estudio.

Durante las últimas dos décadas, el Estado de Florida se ha visto gravemente amenazado por un posible desabastecimiento generado por la reducción del nivel de la superficie potenciométrica de su fuente de agua más importante, el Acuífero Superior Floridano (Clarke et al., 2010). Esta situación ha impulsado a investigadores y entidades gubernamentales a monitorear la calidad del agua de diversos cuerpos a lo largo del Estado, incluida la cuenca del río San Juan Bautista, a la que pertenece el lago George, para evaluar su potencial como fuente alternativa de abastecimiento. La revisión de antecedentes indicó que existe una amplia disponibilidad de estudios desarrollados mediante muestreo y análisis de laboratorio; no obstante, no se hallaron investigaciones que incorporaran técnicas de percepción remota para estimar parámetros de calidad del agua y aportar el monitoreo continuo. Ante este panorama, resultó pertinente presentar y desarrollar una metodología basada en imágenes satelitales recientes que permitiera complementar los enfoques tradicionales y ampliar las posibilidades de seguimiento espacial y temporal con la calidad del agua.

La selección del lago George como zona de estudio también estuvo motivada por la experiencia profesional adquirida durante el desarrollo de proyectos relacionados con el análisis hidrogeológico en el Estado de Florida. Dichas actividades permitieron identificar no solo la existencia de una amplia base de información histórica sobre la calidad del agua del sistema de la cuenca del río San Juan Bautista, sino también un vacío en la aplicación de metodologías basadas en percepción remota para el monitoreo de este tipo de cuerpos hídricos. Adicionalmente, la disponibilidad de una base de datos de mediciones extensa y robusta constituyó un criterio determinante para la selección del área de estudio. Durante la fase de formulación de la investigación se revisó la disponibilidad de datos equivalentes para cuerpos de agua en el territorio colombiano; no obstante, se evidenció que los registros existentes no ofrecían una cobertura temporal ni una densidad de observaciones comparables, lo que se traducía en una limitación notable para el desarrollo y validación de los modelos propuestos. En

consecuencia, el lago George representó un escenario metodológicamente idóneo para evaluar el potencial de las imágenes Sentinel-2 en la estimación de parámetros de calidad del agua y el posterior análisis con un índice de calidad.

La complejidad de determinar una metodología generalizada para evaluar la calidad de cualquier cuerpo de agua ha sido suscitada por el Consejo Nacional de Investigación de Estados Unidos (National Research Council, 2012) al tomar como caso de estudio los diferentes afluentes del Río San Juan Bautista al norte del Estado de Florida, dado que existen diferentes enfoques que pueden llegar a determinar, con distintas escalas y unidades de medida, el nivel de calidad del agua. Dichos enfoques están generalmente desarrollados bajo procedimientos tradicionales de análisis de componentes biogeoquímicos en laboratorio con muestras tomadas *in-situ*. Si bien esta técnica ha mostrado resultados determinísticos y confiables, también demanda una cantidad de tiempo y recursos económicos considerables, aspectos que limitan un monitoreo continuo de los cuerpos hídricos. En consecuencia, la implementación de una estrategia de análisis alternativa por medio de técnicas de percepción remota y aplicación de modelos de *machine learning* permitiría evaluar y supervisar los cambios en la calidad del agua de un cuerpo determinado de una manera más rápida, precisa y confiable.

Como ya se ha venido mencionando, la aplicación de técnicas de inteligencia artificial, específicamente hablando, modelos de *machine learning*, ha venido tomando cada vez más importancia debido a los resultados de los análisis adaptados en distintos fenómenos, por tal motivo, para la evaluación de la estimación de los parámetros de calidad del agua se ha buscado emplear uno de estos modelos, el cual ya ha sido probado en el ámbito de la hidrología demostrando un mejor ajuste sobre otros modelos (Najafzadeh and Basirian, 2023). Dicho modelo es la denominada Regresión con Splines Adaptativos Multivariada (MARS por sus siglas en inglés), el cual no ha sido empleado en la determinación del índice de calidad del agua sobre estudios hechos en la cuenca alta del Río San Juan Bautista, más concretamente, en el lago George. Con la aplicación de este modelo, de igual forma, se buscó confirmar la viabilidad y significancia estadística sobre datos hidrológicos tomados con técnicas de percepción remota en lugares con características hídricas dispares.

La realización de esta propuesta examinó la posibilidad de complementar las investigaciones enfocadas en la generación de parámetros de calidad del agua sobre la zona de estudio y la validación del uso de modelos matemáticos enfocados a la captura de la estructura general de fenómenos de carácter hidrológico. Asimismo, buscó implementar una metodología alternativa a las ya usadas por el Consejo Nacional de Investigación de Estados Unidos y por el Distrito de Gestión de Agua del Río San Juan Bautista para el monitoreo del cambio en la calidad del agua, las cuales están enfocadas en el procesamiento de muestras por vías tradicionales, es decir, evaluaciones de laboratorio, añadiendo técnicas de percepción remota a los resultados finales.

3

El Problema de Investigación

Actualmente, y desde hace unas décadas, el Estado de Florida ha extraído la mayor parte del agua potable del Acuífero Floridano Superior (UFA, por sus siglas en inglés), considerado como la fuente hídrica principal para todo el territorio. Pues bien, el Consejo Nacional de Investigación (RNC, por sus siglas en inglés – Consejo de ahora en adelante-) empezó en el año 2009 la publicación de una serie de reportes enfocados a la investigación de los niveles y calidad de agua del UFA, dado que entre los años 2005 y 2006, el Distrito de Gestión de Agua del Río St. Johns (SJRWMD, por sus siglas en inglés -Distrito de ahora en adelante-) llevó a cabo un conjunto de estudios denominados Estudio de Impacto en el Suministro de Agua (WSIS, por sus siglas en inglés) sobre el considerable aumento en la extracción de líquido, llegando a la conclusión, luego de una minuciosa proyección con datos tomados directamente del acuífero, que para el año 2013 este cuerpo de agua subterránea iba a enfrentar graves problemas de decrecimiento en la superficie potenciométrica superficial (National Research Council, 2012).

La premisa anterior abre la puerta a diferentes cuestionamientos como: ¿Qué alternativas puede haber como suministro de agua potable para el Estado? ¿De qué manera se puede determinar que las fuentes de agua seleccionadas como suministro complementario tengan las características necesarias para el consumo humano? ¿Qué alternativas hay si se encuentra una fuente de agua cuya calidad no sea la esperada? Éstas y más preguntas fueron las que la actual investigación pretendió responder.

Como ya se mencionó anteriormente, el Distrito y el Consejo, dada la preocupación por encontrar una fuente alterna al UFA, empezaron un proyecto enfocado en determinar la viabilidad en cuanto al uso y extracción de agua del Río San Juan Bautista, determinando las consecuencias hidrológicas e hidrodinámicas que este acto conllevaría tanto para la población que fuese a recibir este recurso primario como para el ambiente ecosistémico que usa las aguas del Río como fuente natural. En este proyecto trabajaron más de 80 investigadores, profesionales y consultores de diferentes áreas que, de forma conjunta e independiente, condujeron investigaciones, estudios y evaluaciones en paralelo en la cuenca completa del Río, las cuales, como era de esperarse, no llegaron a una conclusión determinante sobre la viabilidad en la extracción de agua debido a la gran cantidad de incertidumbre a lo largo del cauce y de la variabilidad de consecuencias en las diferentes zonas y territorios que atraviesa el cuerpo.

Este estudio, entre otras recomendaciones y conclusiones, determinó que la cuenca puede seguir siendo material de estudio para futuras decisiones respecto a la búsqueda de una fuente de agua fiable para la población del Estado de Florida. Fue así, entonces, que la presente investigación buscó desarrollar una metodología alterna para el diagnóstico del estado de la calidad del agua en el año 2023 del río San Juan Bautista, específicamente en el lago George, mediante la implementación de técnicas de percepción remota sobre imágenes satelitales, análisis de regresión multivariada y procesamiento con técnicas de *machine learning*. Dada la gran extensión del lago, el proyecto se desarrolló

mediante imágenes satelitales del programa Sentinel-2 para el año 2023 (incluyendo datos de los últimos meses de 2022 y de los primeros de 2024), cuya misión fue lanzada en el año 2015. Estas imágenes fueron fácilmente adquiribles y descargables desde el geoportal de Copérnicus. Adicionalmente, las mediciones con las cuales se generarían y validarían las ecuaciones de los parámetros fueron tomadas del Atlas de Agua del Condado de Lake, el cual cuenta con más de 135.000 muestras de más de 130 parámetros distintos desde el año 1936, solamente en el lago George.

3.1. Planteamiento del Problema.

El uso indiscriminado de los recursos y el crecimiento poblacional son dos de las causas más importantes que se pueden hallar al analizar el problema del decrecimiento en el nivel de la superficie potenciométrica del UFA (Marella and Berndt, 2005). A partir de esta problemática se crea la necesidad de buscar nuevas fuentes de agua para abastecer a los más de 23 millones de habitantes que actualmente viven en el Estado de Florida (Marin, 2024). Dado que la fuente primaria del líquido preciado es subterránea, el Consejo, junto con el Distrito, ha buscado determinar si una alternativa superficial puede ser la solución como complemento a un posible desabastecimiento.

Si bien en un análisis superficial esta puede llegar a ser una idea a considerar, hay varios aspectos a tener en cuenta relacionados con las implicaciones ecológicas y de salubridad que esto tendría; por ejemplo, Florida se caracteriza por tener un ecosistema húmedo y pantanoso, donde diferentes tipos de especies animales y plantas utilizan las aguas del Río San Juan Bautista en el desarrollo de sus ciclos naturales. Por lo tanto, al afectar de alguna forma el nivel del agua del afluente o incluso contaminarlo, podría llegar a tener un impacto ecológico grave, lo cual llevaría a desechar la idea de usar la cuenca, o cualquiera de los cuerpos que la componen, como fuente de agua activa. Por otra parte, están las condiciones de salubridad en cuanto a la potabilidad del agua para consumo humano. La Florida es un territorio en su mayoría plano, con pequeñas variaciones de altura, por lo que los cuerpos de agua que la atraviesan, en su mayoría, cuentan con un caudal bastante bajo; es decir, el agua no circula o lo hace de una manera muy lenta, lo que conlleva la formación de bacterias y microorganismos que podrían afectar el nivel de calidad del agua (Madigan et al., 2010).

En consecuencia, el desarrollo de técnicas más eficientes para el monitoreo de la calidad del agua es una necesidad imperativa debido a que los métodos convencionales, si bien siguen siendo efectivos y determinantes, no dejan de ser un proceso costoso que consume mucho tiempo y recursos. Los resultados de esta investigación correspondieron, entonces, a la implementación de una alternativa metodológica que sirviera como aporte técnico al procesamiento y análisis de datos hidrológicos para la evaluación de la calidad del agua con herramientas no utilizadas o contempladas en estudios anteriores, generando así un argumento adicional para la búsqueda de fuentes de agua alternas que permitan solucionar el problema de la extracción masiva del recurso hídrico con el fin de evitar problemas de desabastecimiento en años futuros.

3.2. Formulación del Problema.

¿Cómo se puede integrar la percepción remota con modelos de *machine learning* para el análisis y estimación del índice de calidad del agua en el Lago George?

3.3. Sistematización del Problema.

- ¿Cuál es la combinación de bandas de imágenes satelitales que mejor estiman los parámetros de calidad del agua?
- ¿Qué tan efectivo es el uso de imágenes satelitales y modelos de *machine learning* en la evaluación del índice de calidad del agua en el Lago George?
- ¿Qué tan significativa es la estimación del índice de calidad del agua mediante la implementación del modelo propuesto?

4

Estado del Arte.

El desarrollo tecnológico de los sensores de las plataformas satelitales ha representado una mejora considerable en la precisión de las investigaciones y trabajos relacionados con el uso de imágenes digitales como insumo principal. Además, gracias a la globalización y distribución de dichos productos de manera gratuita a nivel mundial, es más sencillo, hoy en día, la adquisición, almacenamiento y procesamiento de información espectral, abonando a la producción científica y académica usando la percepción remota como técnica para solucionar una problemática determinada. En este orden de ideas, numerosos estudios han abarcado el uso de imágenes digitales combinadas con técnicas de *machine learning* tanto para la estimación y monitoreo de parámetros como la evaluación de índices de calidad del agua.

Uno de los primeros ejercicios relacionados con imágenes satelitales y técnicas de *machine learning* fue realizado por (Chebud et al., 2012). Esta investigación buscó desarrollar un modelo que predijera parámetros de calidad de agua a lo largo de un área extensa y logísticamente complicada de acceder, como lo es la cuenca del río Kissimmee, al sur del estado de Florida. Las imágenes utilizadas fueron tomadas por el programa Landsat entre dos épocas diferentes; la primera fue entre los años 1998 y 1999, y la restante entre los años 2009 y 2010. El modelo utilizado fue una red neuronal artificial que se aplicó a las imágenes para estimar concentraciones de clorofila-a, turbidez y fósforo antes y después de la restauración ecosistémica y durante temporadas secas y húmedas. La evaluación y validación del modelo demostraron que la red neuronal presentó un gran ajuste en cuanto a la estimación de los parámetros de calidad del agua, con un coeficiente de determinación R^2 mayor al 0.95. Adicionalmente, el error medio cuadrático para el fósforo, turbidez y clorofila-a estuvo por debajo de 0.03 miligramos por litro, 0.5 NTU y 0.17 miligramos por metro cúbicos, respectivamente, probando la utilidad y precisión de las estimaciones en ambientes complejos.

Debido al crecimiento demográfico y agrícola que presentó la bahía costera de Tampa Bay, además del incremento sustancial en el uso de fertilizantes y del paso de los huracanes Frances y Jeanne en 2004 y Wilma en 2005, seguido por una sequía continua de 2006 a 2008 experimentada en su mayoría al sur del estado, la calidad del agua de esta zona del sur y centro occidente de Florida sufrió un gran impacto; por tal motivo, (Chang et al., 2013) desarrollaron un artículo para modelar los patrones espacio-temporales de la concentración de fósforo en dicha zona con la aplicación de modelos de programación genética sobre información espectral extraída de imágenes espectrorradiométricas de media resolución, también conocidas como MODIS. Con la aplicación del modelo, se logró, de igual forma, determinar los impactos inducidos tanto por fuentes naturales como por perturbaciones antropogénicas en los cuerpos de agua cercanos a la zona de estudio.

Adicional al oxígeno disuelto, sólidos suspendidos totales y turbidez, (El Din et al., 2017) desarrolló una metodología para estimar la concentración de parámetros de calidad de agua superficiales

tales como la demanda de oxígeno químico y biológico mediante el uso de imágenes satelitales generadas por el programa Landsat-8 para junio de 2015 y abril y mayo de 2016 en la zona del río San Juan Bautista de la provincia de *New Brunswick* ubicada al suroriente de Canadá. Esta investigación buscó modelar la respuesta espectral de los 5 parámetros anteriormente mencionados con la aplicación de una red neuronal de retropropagación enfocada, justamente, en imágenes Landsat. Uno de los estadísticos utilizados para la validación de las estimaciones fue el coeficiente de determinación R^2, el cual arrojó los siguientes resultados: 0.991, 0.933, 0.937, 0.930 y 0.934 para la turbidez, demanda de oxígeno químico, demanda de oxígeno biológico y oxígeno disuelto, respectivamente.

Análogamente, (Gholizadehm and Melesse, 2017) presentaron un estudio enfocado en el análisis de variabilidad espacio-temporal de parámetros de calidad de agua en la Bahía de Florida. Las variables elegidas fueron la turbidez, la clorofila-a y la concentración de fósforo y nitrógeno totales en tres épocas distintas; la primera, en el año 2000 haciendo uso de imágenes Landsat *Thematic Mapper* (TM); la segunda, en el año 2007 con el mismo programa sensor anteriormente mencionado; y la tercera para el año 2015, esta vez con el sensor *Operational Land Imager* (OLI), igualmente del programa Landsat. El modelo predictivo utilizado para estimar la clorofila-a y la turbidez fue una regresión lineal múltiple *stepwise* arrojando un coeficiente de determinación R^2 de 0.86 y 0.84, respectivamente, en época seca y de 0.66 y 0.63 en época húmeda. Para la concentración de fósforo y nitrógeno totales se utilizó una regresión lineal múltiple *best-fit* arrojando un coeficiente de determinación R^2 de 0.74 y 0.82, respectivamente, en época seca, y de 0.69 y 0.82 en época húmeda.

Un problema que identificaron (Arias-Rodríguez et al., 2021) en cuanto al modelado de parámetros de calidad del agua haciendo uso de técnicas de percepción remota estaba relacionado con la limitada cantidad de información de campo disponible para su uso. En consecuencia, desarrollaron una investigación para complementar el Sistema Nacional Mexicano de Monitoreo de Calidad del Agua (RNMCA) tomando datos levantados entre los años 2013 y 2019 junto con el uso de imágenes provenientes de los programas Landsat-8 OLI, Sentinel-3 OLCI y Sentinel-2 MSI. Adicionalmente, se utilizaron modelos de aprendizaje extremo (ELM), vectores de regresión de apoyo (SVR) y regresiones lineales para la estimación de clorofila-a, turbidez, materia suspendida total y profundidad del disco Secchi. Debido a la poca información de campo con la que se contaba, el ajuste de los modelos no fue tan representativo comparado con los resultados de otras investigaciones. Tanto así que, por ejemplo, la turbidez fue uno de los parámetros con mejor ajuste con un coeficiente de determinación R^2 de 0.7, haciendo uso del sensor OLCI nivel 2 del programa Sentinel-3.

Un enfoque más complejo en cuanto a la estimación y predicción de parámetros de calidad del agua usando modelos de *machine learning* fue desarrollado por (Ahmed et al., 2022b) con el uso de imágenes Landsat-8, las cuales cuentan con una resolución espacial de 30 metros, entre los años 2014 y 2021 en la cuenca Rawal de la ciudad de Islamabad, capital de Pakistán. Los parámetros estimados fueron el oxígeno disuelto y la conductividad eléctrica, teniendo en cuenta variables adicionales como la clorofila-a, la profundidad del disco Secchi, pH, turbidez, sólidos disueltos totales y temperatura superficial. Este estudio empleó un total de 5 modelos de aprendizaje profundo: red neuronal convolucional, red neuronal totalmente conectada, red neuronal recurrente, perceptrón multicapa y memoria de corto a largo plazo, siendo este último el modelo que mejor ajuste presentó con un error medio cuadrático de 0.2 miligramos por litro en la estimación del oxígeno disuelto y de 281.741 microsiemens por centímetro para la estimación de la conductividad eléctrica, demostrando que el enfoque híbrido entre técnicas de monitoreo y estimación tradicionales con respecto al uso de percepción remota, ofrece resultados eficientes y precisos en la predicción de parámetros de calidad de agua.

La cuenca de Rawal, tiempo después, fue objeto de observación por (Ahmed et al., 2022a) para el estudio y desarrollo de un índice de calidad del agua mediante la aplicación de modelos de *machine learning* semisupervisados sobre parámetros estimados de imágenes satelitales. La investigación surgió dada la necesidad de modelar zonas con poca información de campo, haciendo alusión a las grandes limitaciones de los índices al ser evaluados en zonas con características hidrológicas diferentes. Para tal fin, no solamente se utilizaron parámetros hídricos, sino que se incluyó información físico-química, del aire, meteorológica, geológica y topográfica. La metodología implementada seguía 5 pasos esenciales: la selección de parámetros, cálculo de subíndices, asignación de pesos, agregación de subíndices y clasificación. Los modelos utilizados fueron del tipo de árboles de decisión: *LightGBM*, bosque aleatorio, *CatBoost*, *AdaBoost* y *XGBoost*. Siendo el primero y tercero de ellos los que mejor ajuste presentaron con un 99.1 % y 99.3 % de precisión.

Siguiendo con la línea del análisis de calidad del agua por medio de índices integrados, (Najafzadeh and Basirian, 2023) modelaron el índice de calidad del agua bajo la metodología del Consejo Canadiense de Ministros de Medio Ambiente en el río Hudson, ubicado entre las ciudades de Nueva York y Nueva Jersey. Los parámetros elegidos fueron: turbidez, sulfato, sodio, potasio, dureza, flúor, oxígeno disuelto, cloruro, arsénico, alcalinidad, pH, nitrato y magnesio, cuyos datos de campo fueron adquiridos entre marzo y junio de 2021. Las imágenes utilizadas se adquirieron del portal especializado del Servicio Geológico de los Estados Unidos (USGS) del programa Landsat-8 OLI-TIRS. Los parámetros anteriormente mencionados fueron estimados por medio de análisis de regresión lineal múltiple y el índice de calidad de agua fue ajustado, evaluado y validado con 4 modelos de inteligencia artificial: Modelo de Árbol M5, Regresión con Splines Adaptativos Multivariada, Programación de Expresión Genética y Regresión Polinomial Evolutiva. Siendo el segundo de estos el modelo que mejor ajuste presentó.

Finalmente, en el ámbito institucional (Guerrero Buelbas, 2022) constituye un antecedente relevante para la presente investigación. En este trabajo se empleó el análisis de la calidad del agua del lago de Tota con mediciones tomadas desde el año 2014 hasta el año 2019 complementado con el uso de 10 imágenes Sentinel-2. El método elegido para el modelado y estimación de los parámetros fue una red neuronal tipo perceptrón multicapa. Posteriormente, se calculó el índice de calidad del agua del IDEAM, el cual se basa en pesos ponderados para 6 indicadores específicos, y finalmente, se empleó una función de transformación inversa para crear datos para cada clase del índice y se entrenaron modelos basados en máquinas de soporte vectorial y *bagged trees* para la clasificación de la calidad del agua a partir de las concentraciones de los modelos propuestos.

Aunque ambas investigaciones comparten el uso de imágenes Sentinel para el monitoreo de la calidad del agua, presentan diferencias metodológicas importantes. Mientras el estudio del lago de Tota se fundamenta en modelos de aprendizaje automático y en el índice ICA del IDEAM, la presente investigación propone una metodología basada en modelos de regresión lineal ajustados mediante mínimos cuadrados generalizados con autocorrelación (GLSAR), lo que permite abordar explícitamente la dependencia temporal presente en las observaciones. Asimismo, la evaluación de la calidad del agua se realizó mediante el índice del CCME, el cual integra un mayor número de parámetros y ofrece un enfoque diferente para la caracterización del estado del recurso hídrico. Finalmente, en lugar de generar escenarios sintéticos para la clasificación del índice, este trabajo empleó un modelo de regresión de splines adaptativos multivariados (MARS) para analizar la relación entre la respuesta espectral de las imágenes Sentinel-2 y los valores continuos del índice de calidad del agua, aportando una estrategia metodológica distinta para la estimación espacial de este indicador.

5

Marco Teórico.

El presente marco teórico tiene como propósito fundamentar conceptualmente la investigación, a partir de la revisión de estudios previos y teorías relacionadas con el análisis de la calidad del agua por medio de índices ponderados, teniendo como insumo principal los valores espectrales capturados por sensores remotos y la subsecuente relación con métodos predictivos de aprendizaje de máquina como la regresión lineal y el modelo MARS. En esta sección se abordan los principales enfoques, conceptos y antecedentes que permiten comprender el problema de investigación y orientar el análisis de los resultados.

La estructura general de la sección se sustenta en la presentación de los 15 parámetros de calidad de agua considerados inicialmente. Consecuentemente, se desarrolla la conceptualización primigenia que dio origen al ICA del CCME, teniendo en cuenta todos los antecedentes que conllevaron a su estructuración final. Posteriormente, se desglosan las características principales del programa Sentinel-2 y de sus productos asociados. Finalmente, se esboza la fundamentación teórica de la regresión lineal y el modelo MARS, incluyendo los supuestos clásicos y las pruebas de significancia estadística, junto con los métodos de selección de funciones y notación general.

5.1. Parámetros de Calidad del Agua.

El agua es un requisito primario para el desarrollo y subsistencia de la vida en la Tierra. No obstante, es de conocimiento general que la calidad del líquido se ha visto deteriorada a lo largo de los años por distintos fenómenos tanto antropogénicos como naturales (Ahmed et al., 2022c). Tradicionalmente, parámetros físico-químicos se han utilizado para analizar y determinar la variación de las propiedades cualitativas de cuerpos hídricos. Entre los más importantes y frecuentes en artículos científicos y académicos se incluyen: la turbidez, el oxígeno disuelto, la temperatura, la clorofila-a, el pH y una variedad de nutrientes como fósforo, cloruro, nitrógeno y potasio. Para el SJRWMD, la protección y restauración de la calidad del agua es una misión principal. Por tal motivo, el monitoreo de las características físico-químicas superficiales comenzó a ser implementado desde 1980. Desde entonces, los esfuerzos se han expandido y mejorado para la medición y procesamiento de los siguientes parámetros:

5.1.1. Fósforo Total.

El fósforo es un elemento que se puede encontrar en formas orgánicas e inorgánicas que contribuyen al crecimiento de algas y plantas acuáticas. La cantidad de fósforo total contiene tanto componentes disueltos como material particulado y proviene de una muestra de agua sin filtrado previo (SJRWMD, 2024). La unidad de medida de este parámetro suele ser miligramos por litro, la cual

expresa la concentración de una sustancia en un líquido. Para la zona de estudio, los valores de concentración de fósforo total para el 2023 van desde 0.0042 mg/l hasta 0.3585 mg/l.

5.1.2. Nitrógeno Total.

Al igual que el fósforo total, el nitrógeno total es un elemento que puede ser encontrado en varias formas tanto orgánicas como inorgánicas y contribuye a la formación y crecimiento de plantas acuáticas y algas. Este parámetro es calculado con la sumatoria total de la concentración de nitrógeno de *Kjeldahl*, nitrato y nitrito (SJRWMD, 2024). El nitrógeno de *Kjeldahl* es una medida que cuantifica la concentración de nitrógeno orgánico y amoniacal ($NH_3 + NH_{4+}$); esta unidad de medida no es utilizada solamente en el análisis de sustancias disueltas en un líquido, sino en la determinación de la cantidad del elemento en alimentos y suelos. Las concentraciones para el 2023 en el lago van desde 0.0023 mg/l hasta 0.1953 mg/l.

5.1.3. Clorofila-a.

La clorofila-a juega un papel crucial en el análisis de la calidad del agua de cualquier cuerpo hídrico debido a que es el pigmento verde que se encuentra en todas las plantas y facilita la fotosíntesis, además de proveer una estimación de la biomasa aproximada de algas. Los valores de la concentración de clorofila se miden en microgramos por litro y se simboliza así μg/l. Este parámetro debe ser procesado en laboratorio debido a que el producto sufre una degradación natural gracias a la feofitina; por lo tanto, requiere que los valores medidos sean corregidos al sustraer los valores de esta sustancia del resultado final (SJRWMD, 2024). Una alta concentración de clorofila reduce la claridad del agua y limita la cantidad de luz para ecosistemas en aguas no profundas. Los valores de concentración de clorofila para el lago van desde 2.99 μg/l hasta 50.38 μg/l.

5.1.4. Profundidad del Disco Secchi.

La profundidad del disco Secchi (también conocida como transparencia del disco Secchi) es una medida que determina el nivel de transparencia de un líquido, el cual se mide bajo cuadrantes blancos y negros. Una profundidad del disco Secchi mayor representa, asimismo, una mayor claridad de agua. La unidad de medida de este parámetros depende del sistema que se utilice; para la zona de estudio, la unidad de medida es en pies (ft) y va desde 1.31 pies hasta 3.94 pies.

5.1.5. Solidos Suspendidos Totales.

Al igual que la profundidad del disco Secchi, los sólidos suspendidos totales es un parámetro que mide la claridad del agua al medir la cantidad de sustancias y partículas no disueltas en un líquido. Una concentración alta de sólidos suspendidos totales está frecuentemente asociada con una reducida claridad del agua y un color más alto (SJRWMD, 2024). La unidad de medida son miligramos por litro. En la zona de estudio, los valores de esta variable van desde 3.6 mg/l hasta 27 mg/l.

5.1.6. Cloruro.

El cloruro es una forma ionizada del cloro. Este elemento es utilizado para la potabilización del agua, ya que ayuda a la eliminación de microorganismos en el líquido, no obstante, su presencia puede darse en formas contaminantes. Una concentración excesiva de cloruro puede resultar perjudicial

para las plantas y animales (SJRWMD, 2024). La unidad de medida es miligramos por litro y en el lago se pueden hallar concentraciones desde 13.81 mg/l hasta 288.97 mg/l.

5.1.7. Oxígeno Disuelto.

El oxígeno es un elemento requerido por peces y otros organismos acuáticos para su subsistencia, y su concentración en el agua está fuertemente vinculada con la fotosíntesis de las plantas que allí habitan, el movimiento natural del líquido y la temperatura superficial del cuerpo objeto de estudio. Hay diversas condiciones naturales que pueden afectar los niveles de oxígeno como aguas subterráneas de manantiales, corrientes de humedales y pantanos, temperaturas considerablemente altas y climas templados o nubosos. El método de medición de oxígeno en el lago se hace por medio de una sonda, en lugar de recolección de muestras tradicional (SJRWMD, 2024). La unidad de medida de este parámetro es miligramos por litro y para el año 2023 se tomaron valores entre 0.35 mg/l y 9.8 mg/l.

5.1.8. Nivel de Acidez (pH).

El nivel de acidez de una sustancia se determina mediante el uso del pH, el cual es una escala que toma valores de 0 a 14. Un valor mejor a 7 significa que el líquido o sustancia es ácida; mientras que, un número superior a 7 indica que la sustancia es básica o alcalina. En consecuencia, un valor de 7 en el pH sugiere una acidez neutra. Los cambios en la acidez o pH de un cuerpo de agua pueden darse por distintos motivos, entre los más representativos están: deposición atmosférica o también llamada lluvia ácida, factores geológicos, vegetación y polución (SJRWMD, 2024). La unidad de medida es adimensional, usando la escala anteriormente mencionada como referencia y para el cuerpo de agua objeto de estudio se encontraron valores entre 6.66 y 8.95.

5.1.9. Temperatura superficial.

El estado de Florida goza de una localización particular en el país que lo diferencia de sus pares en cuanto a estaciones o temporadas climáticas. Al encontrarse más cerca al trópico y gracias a la escasa variación de altitud, el clima está más inclinado a un sistema bimodal que estacional, es decir, existen temporadas secos y húmedas más allá de las conocidas invierno, otoño, verano y primavera (Lund and Curry, 2006). Este fenómeno incide directamente en el desarrollo y tasa de crecimiento de plantas y animales en los cuerpos de agua. Al igual que con la profundidad del disco Secchi, la unidad de medida depende del sistema que se use, en este caso, la temperatura se mide en grados Fahrenheit con valores que van desde los 56.12 °F hasta 89.13 °F.

5.1.10. Turbidez.

La turbidez es una medida que representa la cantidad de luz que penetra un cuerpo de agua. Este parámetro está directamente relacionado con la claridad del agua y puede ser fuertemente influenciado por la cantidad de partículas suspendidas en la superficie. Estas partículas pueden surgir a causa de la erosión de la cuenca que transporta componentes del suelo y materia orgánica. El tránsito de botes y vehículos acuáticos en general puede llegar a agitar sedimentos y causar una turbidez más alta. Otra causa son las altas concentración de nutrientes que pueden ocasionar un crecimiento excesivo de algas (LCWA, 2025). La unidad de medida es NTU que significa *nephelometric turbidity unit* o unidad de turbidez nefelométrica la cual indica, justamente, la cantidad de partículas suspendidas en un fluido. Teniendo valores altos de NTU, significa que el líquido se ve más sucio en apariencia.

Según las mediciones del lago para el 2023, los niveles de turbidez van desde 0.17 NTU hasta 12.28 NTU.

5.1.11. Arsénico.

El arsénico es un elemento natural presente en la corteza terrestre distribuido en el medio ambiente. Se puede encontrar en el aire, el agua y la tierra. Una particularidad de este elemento es que encontrado en su forma inorgánica es muy tóxico. No obstante, al igual que con el cloruro, niveles de concentración bajos se pueden hallar en agua potable sin representar un peligro tanto para animales y plantas como para el consumo humano (WHO, 2022). La Organización Mundial de la Salud establece un límite de 10 μg/l para considerar un nivel máximo de arsénico. Las regiones en las que es más probable encontrar este mineral son zonas con actividad volcánica o minera activas, sin embargo, este parámetro también puede darse de manera natural. Como ya se mencionó anteriormente, la unidad de medida de la concentración de arsénico es microgramos por litro y para la zona de estudio se hallaron valores de 0.046 μg/l hasta 0.81 μg/l.

5.1.12. Dureza.

La dureza hace referencia a la cantidad de calcio y sales de magnesio en el agua. Si bien otros cationes pueden contribuir a la concentración total de dureza, estos dos compuestos son los más comunes y se presentan en pequeñas cantidades en agua dulce. Comercialmente, se encuentra en el mercado distinta clase de equipos que mide la dureza en cantidades de calcio carbonatado ($CaCO_3$) en un líquido o fluido determinado (Astrofsky and Sagerstrom, 2002). La unidad de medida más común para este parámetro es partes por millón (ppm) o, miligramos por litro (mg/l). En términos de esta última medida se encuentran las concentraciones para el lago George que van desde 103.30 mg/l hasta 251.70 mg/l.

5.1.13. Potasio.

El potasio es un catión que se halla en el agua, en su mayoría, debido a procesos naturales como la disolución de minerales o por factores antropogénicos como la actividad agrícola y la industria. Las concentraciones comúnmente encontradas no representan un peligro ni para peces y plantas ni para el consumo del ser humano; no obstante, aquellas muy elevadas implican un desarrollo más limitado de la vida ecosistémica, además de tener implicaciones para la salud de las personas (NIH, 2019). Las concentraciones se miden en miligramos por litro, y en el lago se encuentran valores desde 0.81 mg/l hasta 5.95 mg/l.

5.1.14. Sodio.

El sodio es un mineral que se puede hallar de forma natural en el agua que proviene de la disolución de este elemento en las rocas y el suelo. También pueden presentarse concentraciones de sodio gracias a actividades agrícolas e industriales. Al igual que el potasio, concentraciones muy altas de este mineral pueden limitar el desarrollo de especies animales y plantas, así como representar una amenaza para la salud (Salas-Salvadó et al., 2021). La Organización Mundial de la Salud ha establecido que valores superiores a 12 mg/l pueden alterar las propiedades del agua, haciéndola menos aceptable para su consumo (WHO, 2022). La unidad de medida, como ya se mencionó, es miligramos por litro y en la zona de estudio se presentan valores desde 8.19 mg/l hasta 165.95 mg/l.

5.1.15. Sulfato.

El azufre es un mineral presente en el agua esencial para los seres vivos. Comúnmente se mide en sulfatos (SO_4) o en sulfuro de hidrógeno (H_2S). Los sulfatos son sales derivadas del ácido sulfúrico (H_2SO_4) que se disuelven fácilmente en el agua. La presencia de este mineral se da tanto de manera natural como por actividad agrícola o industrial. Por otra parte, el sulfuro de hidrógeno es un gas que se disuelve en el agua y es el responsable de generar un olor fuerte en el agua debido a la descomposición de materia orgánica en condiciones anaeróbicas (INSST, 2017). Las concentraciones de Sulfato se miden en miligramos por litro y en la zona de estudio se pueden encontrar valores desde 0.9 mg/l hasta 83.44 mg/l.

5.2. Índice de Calidad del Agua.

Los índices de calidad del agua tradicionales consistían en el análisis estadístico variable por variable o cuerpo de agua por cuerpo de agua, lo que hacía del proceso una tarea tediosa y no extrapolable a diversos entornos debido a la especificidad de los datos y procedimientos que se debían llevar a cabo. Así entonces, el Consejo Canadiense de Ministros de Medio Ambiente (CCME, por sus siglas en inglés) desarrolló una nueva metodología que buscaba simplificar los estudios en un solo índice, el cual denominaron el índice de calidad del agua del CCME (WQI CCME) (Neary et al., 2001). Cabe aclarar que esta metodología no fue elaborada con el fin de unificar todos los esfuerzos en el cálculo de una sola variable, sino que su propósito es ofrecer una alternativa adicional que permita obtener una idea general de la situación real del cuerpo de agua objeto de estudio.

Entre las ventajas más importantes se encuentran: su capacidad para representar mediciones de diferentes tipos en un solo índice; la capacidad de combinar valores en diferentes unidades de medida en una única métrica adimensional y la facilidad en la comunicación de los resultados. No obstante, no todo son ventajas; debido a unificación de parámetros, el índice tiende a la pérdida de información, por ejemplo, en cuanto a la relación entre variables. De igual forma, dependiendo de la formulación del índice, los resultados pueden variar lo que conduce a una interpretación errada de éste. Asimismo, el CCME WQI tampoco fue creado para reemplazar los métodos de evaluación convencionales de la calidad del agua como la medición de muestras en campo o la utilización de otros índices, es más, la aplicación de esta técnica presupone una adquisición y análisis de los parámetros adecuados.

5.2.1. Reseña Histórica de los Índices de Calidad de Agua Predecesores al CCME WQI.

Hasta la creación y divulgación del CCME WQI, en Canadá existían diferentes instituciones y jurisdicciones que utilizaban distintos tipos de unidades métricas para analizar la calidad del agua. En consecuencia, en el año 1997 el CCME creó el subcomité Técnico del Índice de Calidad del Agua con el fin de evaluar diversos enfoques para la formulación de una metodología más simple y directa que facilitara la diseminación y reportaje de estudios en el país. A continuación, se describen los índices que se usaban antes de su unificación:

5.2.2. Centro de St Laurent.

El Centro *St Laurent* (CSL) era una oficina medioambiental canadiense responsable de reportar el monitoreo de la calidad del agua en el río *St. Lawrence* el cual seguía la siguiente estructura:

$$WQI_{CSL} = \frac{\sum(A_i * F_i)}{n} \tag{5.1}$$

Donde A_i es el promedio del excedente por cada variable i para cada nivel guía i. Cuando el valor de una variable excedía el valor guía para una variable específica, la tasa excedente-valor guía se calculaba. Luego, estas tasas se sumaban y se dividían por el número n de ocasiones en las que ese suceso pasaba. F_i es la frecuencia en la que los valores excedían el valor guía para una variable relativa al total de valores medidos para esa misma variable, es decir:

$$F_i = \frac{F_{excedida}}{F_{total}}, \text{ para toda variable } i \tag{5.2}$$

El CSL calculaba distintos tipos de índices dependiendo del uso que se le fuese a dar al cuerpo de agua. Un índice diferente era calculado para lineamientos de protección de vida acuática, otro para recreación, uno diferente para consumo humano y así sucesivamente.

5.2.3. Quebec.

El índice de Quebec (Hébert, 1996), a su vez, estaba basado en otro índice desarrollado en Nueva Zelanda (Smith, 1990), el cual consistía en la determinación del valor mínimo dentro de una serie de subíndices que se calculaban para cada parámetros de calidad de agua medido, representado de la siguiente manera:

$$WQI = \min\left(I\text{sub}_1, I\text{sub}_2, \ldots, I\text{sub}_n\right) \tag{5.3}$$

Lo que hacía diferente a este índice era que hacía uso de curvas de *Delphi* para el cálculo de los subíndices. Estas curvas se basaban en la opinión de expertos en cuanto a la significancia de un nivel particular de un componente de calidad de agua determinado, es decir, representan la opinión agregada de lo que significa un nivel particular de un contaminante específico en términos del uso designado del agua.

5.2.4. British Columbia.

El índice de calidad del agua de *British Columbia* (Rocchini and Swain, 1995) incluía un factor que no se había tenido en cuenta antes y se representaba con la siguiente estructura:

$$WQI_{BC} = \left(F_1^2 + F_2^2 + F_3^2\right)^{\frac{1}{2}} \tag{5.4}$$

Donde F_1 es el porcentaje de valores guía de calidad del agua excedidos; F_2 es el porcentaje de medidas en las que una o más valores guía fueron excedidos y F_3 es el máximo (normalizado a una escala de 100) por el cual se excedió el valor guía.

Este índice implicaba una aplicación y desarrollo más extenso y tedioso dentro de todos los que se utilizaban en la época; dos de los factores son análogos a componentes de otros índices F_2 es bastante similar al índice de Alberta, mientras que F_1 consistía en un cálculo parecido al del índice del CLS. La provincia de Manitoba adoptó este mismo índice, concluyendo que los resultados obtenidos demostraron ser razonables, por lo que no se propuso ninguna modificación.

5.2.5. Alberta.

Alberta utilizó un indicador de rendimiento para la calidad del agua representado de la siguiente manera:

$$WQI_A = \left(\frac{n_{excedido}}{n_{medido}}\right) * 100 \tag{5.5}$$

Lo que representaba este índice era el porcentaje de muestras tomadas *in-situ* cuya medición excedía los valores guía para un uso designado dentro de los cuales se incluían: recreacional, agrícola y vida acuática.

5.2.6. Ontario.

El índice que la provincia de Ontario usó fue el de *British Columbia*, no obstante, se le modificó el factor F_3 el cual representaba el promedio normalizado de los excedentes en lugar del valor máximo. Este cambio se hizo debido a que se reportó que el tercer factor de dicho índice se saturaba cuando los valores de excedencia eran demasiado amplios.

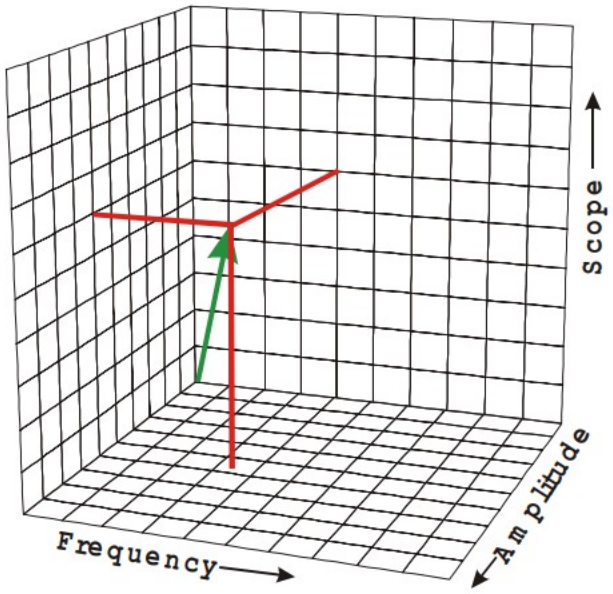


FIGURA 5.1: Modelo conceptual del CCME WQI.

Fuente: (Neary et al., 2001).

5.2.7. Modelo Conceptual del CCME WQI.

El subcomité, en primera instancia, adoptó el enfoque del indice de *British Columbia*, sin embargo, se descartó debido a la logística, tiempo y recursos económicos requeridos para implementar las curvas de *Delphi* cuando se manipulaba un gran número de variables. Ante este panorama, se consideró otro planteamiento. el cual incluye 3 factores escalados de 0 a 100. Los valores de las 3 medidas simbolizan un objetivo de superación el cual se proyecta en un espacio tridimensional donde se combinan los valores de los vectores en un mismo punto, tal como lo muestra la figura (5.1). Este índice toma valores de 0 a 100, por lo tanto, los valores más cercanos a 0 se traducen en una calidad de agua pobre, en consecuencia, valores cercanos a 100 representan una calidad bastante buena.

Una vez completada la primera versión del índice, el subcomité evaluó el comportamiento y rendimiento usando datos sintéticos (Hart, 1998) y conjuntos de datos de *British Columbia* (Phippen, 1998) y *Newfoundland* (Husain, 1998). Al igual que con datos de Alberta y Ontario. Esta evaluación reveló que la estimación de la frecuencia y la amplitud generaba problemas significativos; por lo tanto, el Departamento Rural y de Alimentación Agrícola de Alberta, en asociación con el Instituto de Protección Medio Ambiental de Alberta, apoyó una reformulación de los factores de alcance y frecuencia, obteniendo como resultado el índice de calidad de agua final.

5.2.8. Formulación Final del CCME WQI.

El CCME WQI consiste en 3 factores esenciales:

Factor 1: Alcance

F_1 (alcance) representa la extensión de los parámetros que no cumplen con los valores guía durante el periodo de tiempo de interés. Factor adoptado directamente del índice de *British Columbia*:

$$F_1 = \left(\frac{N_f}{T_v}\right) * 100 \tag{5.6}$$

Donde N_f es la cantidad de variables que fallaron y T_v el total de variables que se incluyeron en el estudio.

Factor 2: Frecuencia

F_2 (Frecuencia) representa el porcentaje de pruebas individuales que no cumplieron con los valores guía, en otras palabras, pruebas falladas, expresado de la siguiente manera:

$$F_2 = \left(\frac{N_t}{T_t}\right) * 100 \tag{5.7}$$

Donde N_t es la cantidad de pruebas que fallaron y T_f el total de pruebas que se llevaron a cabo. Este factor también se derivó directamente del índice de *British Columbia*.

Factor 3: Amplitud

F_3 (Amplitud) representa la cantidad en la que los valores de prueba fallidos no cumplen con los objetivos. Este factor se calcula en tres pasos.

- El número de veces en que una concentración individual es más mayor (o menor cuando el objetivo o valor guía es mínimo) que el objetivo se denomina excursión y se expresa:

$$Ex_i = \left(\frac{FTV_i}{Obj_j}\right) - 1 \tag{5.8}$$

Donde FTV_i es el valor de la prueba fallida y Obj_j el valor objetivo que no se cumplió. Para los casos en los cuales el valor de prueba no debe ser menor al objetivo, la ecuación se reformula de la siguiente manera:

$$Ex_i = \left(\frac{Obj_j}{FTV_i}\right) - 1 \tag{5.9}$$

- La cantidad colectiva de pruebas que no cumplen se calcula al sumar todas las excursiones de las pruebas individuales con respecto a sus valores objetivo y dividiendo ese valor por el número total de pruebas realizadas. Esta variable se denomina *nse* y se calcula así:

$$nse = \frac{\sum_{i=1}^{n} Ex_i}{NumPruebas} \tag{5.10}$$

- El último factor F_3 es calculado entonces por una función asintótica que escala la suma normalizada de las excursiones de los objetivos para obtener un rango de 0 a 100, representado de la siguiente manera:

$$F_3 = \left(\frac{nse}{0{,}01 * nse + 0{,}01}\right) \tag{5.11}$$

Entonces, el CCME WQI se calcula así:

$$WQI = 100 - \left(\frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1{,}732}\right) \tag{5.12}$$

El factor de 1,732 surge debido a que cada factor individual puede tomar un valor máximo de 100 por lo que la raíz resultaría en un valor máximo de 173,2 como lo muestra la ecuación (5.13); por lo tanto, al dividir por este valor, el índice vuelve a escalarse de 0 a 100:

$$\sqrt{100^2 + 100^2 + 100^2} = \sqrt{30000} = 173{,}2 \tag{5.13}$$

5.2.9. Categorización del CCME WQI.

La asignación de los valores del CCME WQI se denomina categorización y representa un proceso crítico y subjetivo a la vez. La categorización debe hacerse con base en la mejor selección de información disponible, criterio experto y las expectativas generales de la calidad del agua. Debido a la naturaleza misma del índice, es muy complejo determinar con base en un rango si la categoría se debe a las excursiones extremas en uno o varios parámetros; no obstante, la estructuración de la categorización está abierta al usuario dependiendo del comportamiento del índice en la zona de estudio. Por consiguiente, una vez se calcula el CCME WQI, se puede categorizar en uno de los siguientes niveles:

TABLA 5.1: Categorías del índice de calidad del agua según el CCME WQI.

Categoría	Descripción
Excelente	(CCME WQI 95–100) Calidad del agua protegida casi sin amenazas; condiciones cercanas a lo natural. Solo ocurre si todas las mediciones cumplen los valores objetivo siempre.
Buena	(CCME WQI 80–94) Calidad protegida con pocas amenazas; condiciones próximas a lo natural o lo deseado.
Aceptable	(CCME WQI 65–79) Calidad usualmente protegida; ocasionalmente amenazada. Condiciones alejadas de lo natural.
Mala	(CCME WQI 45–64) Calidad frecuentemente amenazada; condiciones muy alejadas de lo natural o de los valores deseados.
Muy mala	(CCME WQI 0–44) Calidad siempre amenazada; condiciones no naturales y muy alejadas de los niveles deseados.

Fuente: (CCME, 2017).

5.3. Programa Sentinel-2.

El programa Sentinel-2 nace de una iniciativa denominada Monitoreo Global para el Medio Ambiente y la Seguridad (GMES) creada por la Comisión Europea (EC) y la Agencia Espacial Europea (ESA). El propósito de la GMES era la adquisición de información observable para su aplicación en el ámbito ecológico y de seguridad. La misión del Sentinel-2 fue lanzada en dos fases; la primera el 23 de junio de 2015 y se denominó Sentinel-2A y la segunda, fue lanzada en el primer trimestre de 2017, denominada Sentinel-2B. Su objetivo principal es el de proveer una continuidad en los servicios basados en observaciones de alta resolución espacial de carácter multiespectral y actualmente se usa en distintos ámbitos como: gestión del riesgo, en el monitoreo de inundaciones e incendios ambientales; cambios del uso de suelo; sistemas de emergencia; manejo del agua y protección del suelo, además de la cartografía urbana y rural (Spoto et al., 2012). El diseño del satélite busca complementar las observaciones de la tierra hechas por los programas hermanos de Landsat (programa estadounidense) y SPOT (programa francés).

5.3.1. Características Generales.

El programa consiste en la operación de dos satélites idénticos trabajando simultáneamente y separados 180° entre sí. La órbita es heliosincrónica y se encuentra a una altitud promedio de 786 kilómetros, con un nodo descendente a las 10:30 am. Este tiempo local fue dispuesto para minimizar la presencia de nubes asegurando una iluminación solar viable. Los satélites adquieren imágenes de la tierra entre los 56° sur hasta los 84° norte. A su vez, la misión se compone de dos segmentos esenciales: el segmento espacial, el cual está compuesto por los dos satélites mencionados anteriormente incluyendo el sensor y todo el equipo; luego está el segmento en tierra el cual se encarga de facilitar la adquisición y divulgación de los productos así como el control de la misión y los equipos en el espacio.

5.3.2. Características del Sensor.

El sensor multiespectral se basa en el concepto de barrido de empuje e incorpora un telescopio del tipo anastigmat de tres espejos. El diámetro de la pupila es de 150 milímetros aproximadamente, la cual se encuentra montada de manera isostática en la plataforma, buscando así minimizar las distorsiones termoelásticas; adicionalmente, el telescopio y los espejos están hechos a base de carburo de silicio, lo que le permite, igualmente, minimizar distorsiones de este tipo. El campo de visión del

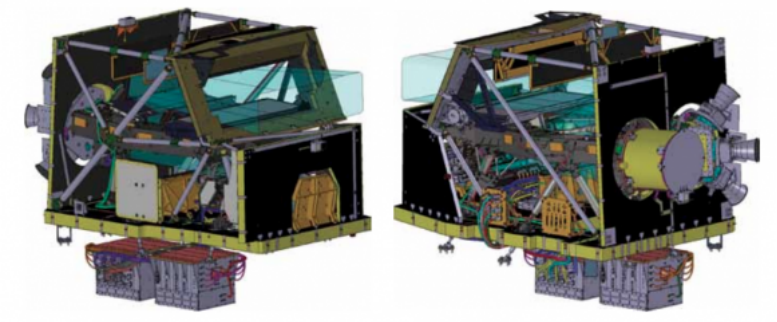


FIGURA 5.2: Características técnicas del sensor.
Fuente: (ESA, 2015).

satélite puede alcanzar los 290 kilómetros, gracias a la optimización y mejora del sensor a lo largo de los años.

El sensor cuenta con 13 canales o bandas espectrales que abarcan desde el espectro visible, pasando por el infrarrojo cercano (NIR) hasta llegar al infrarrojo de onda corta (SWIR) como lo muestra la figura (5.3). La longitud de onda de cada banda se resume de la siguiente forma:

TABLA 5.2: Bandas del programa Sentinel-2: Longitudes de onda y resolución espacial.

Banda	Lon. de onda (nm)	Lon. de onda media (nm)	Resolución (m)
1 - Aerosol	430-450	442,7	60
2 - Azul	450-520	492,7	10
3 - Verde	540-570	559,8	10
4 - Rojo	650-680	664,6	10
5 - Infrarrojo 1	690-710	704,1	20
6 - Infrarrojo 2	730-740	740,5	20
7 - Infrarrojo 3	770-790	782,8	20
8 - Infrarrojo cercano 1	780-900	832,8	10
8A - Infrarrojo cercano 2	850-870	864,7	20
9 - Vapor de agua	930-950	945,1	60
10 - Cirrus	1360-1390	1375,0	60
11 - Infrarrojo de onda corta 1	1560-1650	1613,7	20
12 - Infrarrojo de onda corta 2	2100-2280	2202,4	20

Fuente: (ESA, 2015).

La resolución espacial puede variar dependiendo de la combinación de bandas que se utilice, iniciando desde los 60 metros hasta los 10 metros de información en su unidad mínima de información. Para asegurar la compatibilidad con el programa SPOT se usan 4 bandas con la resolución de 10 metros. Un adicional de 6 bandas cuenta con una resolución de 20 metros para los productos de nivel 2 y que se relacionan directamente con el monitoreo de la vegetación. Las bandas restantes están enfocadas a la corrección atmosférica y cribado de nubes; esta resolución está más relacionada con el estudio y captura de la variabilidad espacial de los parámetros geofísicos y atmosféricos de la Tierra (Spoto et al., 2012).

5.3.3. Tipos de Productos.

Los datos provenientes del satélite deben calibrarse tanto geométricamente como radiométricamente. Los productos de Nivel 0 y 1A se ofrecen brutos, es decir, sin ningún tipo de corrección. Mientras que los datos Nivel 1B son radiancias a las que se les aplica un proceso de corrección radiométrica previa. En cuanto a la corrección geométrica, el modelo físico geométrico se ajusta y refina con puntos de control terrestre y se adjunta al producto, pero no se aplica. Existe una variedad de productos adicionales denominados Nivel 1C, los cuales proporcionan reflectancia geocodificada del denominado tope de atmósfera, el cual incluye registros multiespectrales y multitemporales de píxeles. El producto final unitario es un mosaico de 100 por 100 kilómetros cuadrados, los cuales son ofertados y puestos a disposición para cubrir el área de interés. Los productos finales del programa Sentinel-2 siguen las recomendaciones emitidas en la Garantía de Calidad para la Observación de la Tierra (QA4EO) por el Comité de Satélites de Observación de la Tierra (CEOS).

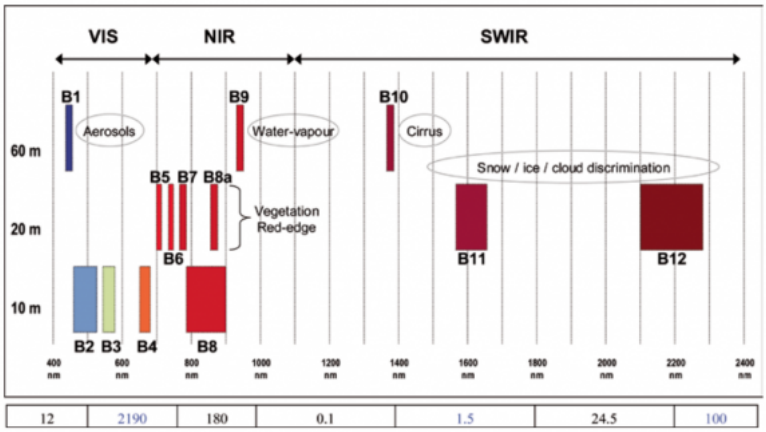


FIGURA 5.3: Cobertura espectral del Sentinel-2.
Fuente: (ESA, 2015).

5.3.4. Nomenclatura de los Productos.

Las imágenes que se obtienen del programa Sentinel-2 se codifican con una nomenclatura específica que permite identificar información relevante tanto de la fecha de toma como de la ubicación geográfica y datos de posprocesamiento del producto. Por ejemplo, una imagen utilizada en la presente investigación del estado de Florida, tomada el 7 de abril de 2024, se identifica de la siguiente forma:

- S2B_MSIL2A_20240407T155819_N0510_R097_T17RMN_20240407T202406.SAFE

Del ejemplo anterior se puede identificar la siguiente estructura:

- MMM_CCCCCC_yyyymmddThhmmss_NPPPP_ROOO_ttttt_YYYYMMDDTHHMMSS.SAFE

Dicha estructura contiene una serie de elementos codificados que identifican la imagen en su fecha de adquisicion y procesamiento, ubicación de la zona capturada y datos relacionados con el sensor y versión del procesador como se evidencia en la tabla (5.3):

TABLA 5.3: Nomenclatura del producto del programa Sentinel-2.

Segmento	Significado
MMM	Satélite que tomó la imagen, puede ser S2A o S2B
CCCCC	Tipo y nivel de procesamiento: Mission Specific Instrument - L2A.
yyyymmddThhmmss	Fecha y hora de adquisición: 7 de abril de 2024 a las 15:58:19 UTC
NPPPP	Versión del procesador: 05.10
ROOO	Órbita relativa: 097
ttttt	Tesela de la malla UTM: zona 17R, cuadrícula MN
YYYYMMDDTHHMMSS	Fecha y hora de procesamiento: 7 de abril de 2024 a las 20:24:06 UTC

Fuente: (ESA, 2026).

5.4. Regresión Lineal Múltiple por MCO.

En términos prácticos, el proceso de regresión lineal es el estudio de la dependencia de una variable con respecto a una o más variables, con el propósito de predecir (o estimar) el valor promedio poblacional de la primera variable, teniendo en cuenta los valores conocidos de las demás (Gujarati and Porter, 2010). El término *regresión* fue utilizado por primera vez por (Galton, 1886) en el marco de una investigación en la que se analizaba el comportamiento de la altura de los hijos con respecto a los padres, donde básicamente se notaba que los hijos de personas altas eran tales, pero no tan altos como sus progenitores; asimismo, los hijos de personas de baja estatura eran tales, pero no tan bajos como sus padres. Por lo tanto, *Francis Galton* concluyó que la altura de las personas tendía a *regresar* o desplazarse hacia un promedio poblacional determinado. Este análisis se resume en la generación de una función en la cual la variable dependiente, también llamada variable explicada, regresada, respuesta, endógena o controlada, está representada por una combinación de variables independientes, también llamadas explicativas, regresoras, estímulo, exógenas o de control. Dicha función sigue la forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \tag{5.14}$$

Donde Y es la variable dependiente; X_1, X_2, ..., X_k son las variables independientes que van a explicar la variable dependiente; u es el término de perturbación estocástico o error; i es la i-ésima observación o dato; β_0 es el intercepto y β_1, β_2, ..., β_k, también llamados estimadores, son el peso de cada variable regresora sobre el modelo. El método de obtención de los estimadores es por medio de mínimos cuadrados ordinarios (MCO) en el cual se busca que los valores de los betas se seleccionen en el caso en que la suma de cuadrados de los residuos sea la más pequeña posible, en otras palabras, que la distancia entre los residuos y el valor promedio esperado sea menor. Es decir:

$$\min \sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_0 - \beta_2 X_{2i} - \cdots - \beta_k X_{ki})^2 \tag{5.15}$$

La solución de la ecuación (5.15) permite obtener la ecuación que determina los estimadores con el menor grado de dispersión de los residuos.

5.4.1. Coeficiente de Determinación Múltiple.

El coeficiente de determinación se simboliza con R^2 y representa el grado de bondad de ajuste de la ecuación de regresión; es decir, da la proporción o porcentaje de la variación total en la variable dependiente Y por las variables explicativas X (Gujarati and Porter, 2010).

$$R^2 = 1 - \frac{\sum \hat{u}_i^2 / (n-k)}{\sum \hat{y}_i^2 / (n-1)} \tag{5.16}$$

Donde, $\sum y_i^2$ es la suma de cuadrados de las observaciones; $\sum \hat{u}_i^2$ es la suma de cuadrados de los residuos; n es el número de observaciones y k el número de parámetros o estimadores.

La estimación de parámetros por medio del método de mínimos cuadrados ordinarios requiere del cumplimiento de ciertos supuestos para poder asegurar que las predicciones realizadas a partir de la ecuación de regresión sean válidas y reflejen de manera correcta y significativa el problema que se está intentando modelar.

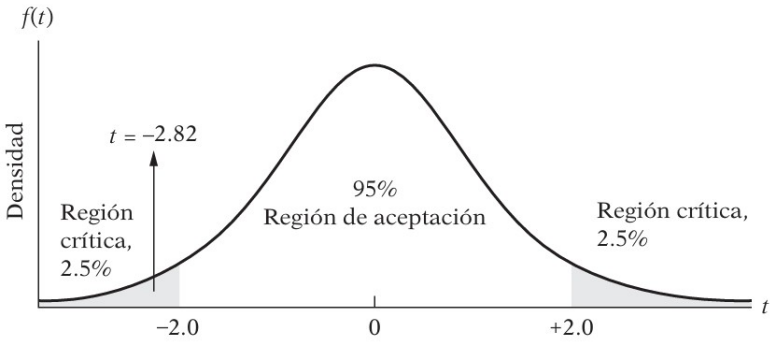


FIGURA 5.4: Prueba de hipótesis de normalidad.
Fuente: (Gujarati and Porter, 2010).

5.4.2. Supuestos de Normalidad.

La regresión lineal múltiple es una técnica de estimación de parámetros, los cuales representan el peso de la variable exógena en el modelo. Pues bien, dicho peso debe asegurar las propiedades de cualquier estimador, las cuales son: consistencia, sesgo y eficiencia. Un estimador es consistente en la medida en que converge al valor esperado al aumentar el número de la población; asimismo, es insesgado en la medida en que el valor esperado es igual al valor verdadero y es eficiente en la medida en que su varianza sea la mejor en comparación con otros estimadores. Al cumplimiento de estas propiedades se le conoce como MELI o mejor estimador lineal e insesgado, lo cual asegura la integridad de las predicciones.

Una forma de corroborar que los estimadores son MELI es por medio del análisis de normalidad de los residuos, los cuales, desde un punto de vista teórico, se pretende que sigan una distribución normal; de esta forma se asegura de que la varianza es constante y tiene media 0. Por lo tanto, el primer, segundo y tercer momento o promedio, varianza y coeficiente de asimetría se requiere que sean o estén muy cercanos a 0; para el cuarto momento o curtosis se requiere que sea o esté muy cercano a 3. Adicionalmente, existen pruebas de hipótesis que ayudan a validar la normalidad o no normalidad de los residuos. Dicha prueba sigue la estructura:

$$\begin{aligned} H_o &: \text{Los residuos siguen una distribución normal} \\ H_a &: \text{Los residuos no siguen una distribución normal} \end{aligned} \tag{5.17}$$

Teniendo un nivel de significancia α, por ejemplo, del 5 %, se puede rechazar la hipótesis nula en el paso en que el valor P sea menor a α. Actualmente, las pruebas de hipótesis más conocidas son: *Shapiro-Wilk*, *Jarque-Bera*, *Kolmogorov-Smirnov*, *Cramer-von Mises* y *Anderson-Darling*

5.4.3. Supuesto de Homocedasticidad.

El término homocedasticidad viene de igual (homo) y dispersión (cedasticidad), es decir, igual varianza. Explicado de otra forma, significa que las poblaciones Y correspondientes a diversos valores de X tienen la misma varianza (Gujarati and Porter, 2010), como lo indica la figura (5.5). Este supuesto indica que la varianza del término de error u_i es la misma sin importar el valor de X, es decir:

$$\text{var}(u_i) = \text{E}[u_i - \text{E}(u_i|X_i)]^2 = \sigma^2 \tag{5.18}$$

Este supuesto también puede ser verificado por medio de pruebas de hipótesis, dentro de las cuales

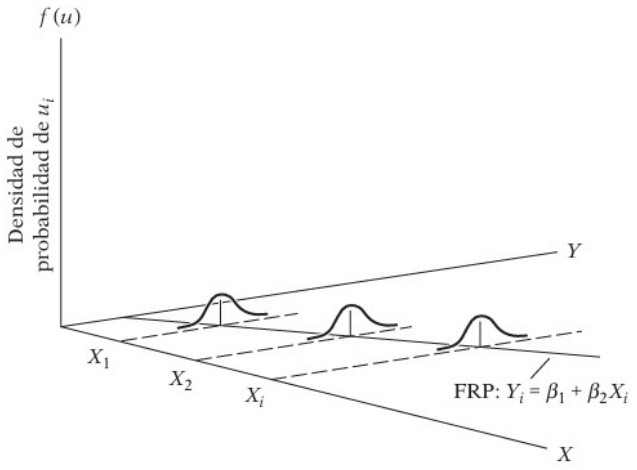


FIGURA 5.5: Comportamiento de homocedasticidad.
Fuente: (Gujarati and Porter, 2010).

se encuentran las pruebas de *Breusch-Pagan*, *White* y *Goldfeld-Quandt*. La prueba tiene la siguiente estructura:

$$\begin{aligned} H_o &: \text{Hay homocedasticidad} \\ H_a &: \text{Hay heterocedasticidad} \end{aligned} \tag{5.19}$$

5.4.4. Supuesto de No Autocorrelación.

Este supuesto indica que los residuos del modelo no están correlacionados, es decir, u_i y u_j son independientes y no tienen relación alguna. Esto significa que las desviaciones de dos valores de Y respecto a sus valores promedio no muestran patrones, como lo indica la figura (5.6). En el caso de que se identifique este patrón determinado en las estimaciones, se confirmará que existe correlación serial o autocorrelación. Un ejemplo gráfico de residuos no autocorrelacionados se ilustra en la figura (5.7). La forma de corregir este supuesto requiere el análisis y determinación de los patrones de correlación, por ejemplo, si un residuo u_i está relacionado con un residuo u_{i-1}, esto quiere decir que la variable Y no solo estaría siendo explicada por X, sino por u_i o u_{i-1} dependiendo del grado de correlación. En este caso, se tendría que contemplar el uso de otro tipo de modelo, como los modelos autorregresivos en el error.

5.4.5. Supuesto de Multicolinealidad.

Este principio indica que las variables explicativas utilizadas para ajustar el modelo tienen algún tipo de correlación entre sí. Esto provoca que la estimación de los parámetros disminuya y que su interpretación sea errónea. Este fenómeno puede surgir por distintas razones, por ejemplo, cuando una variable es una combinación lineal de otras o cuando dos variables expresan casi la misma información (Pérez and Fernández, 2008). A diferencia de los supuestos anteriores, no existe una prueba de hipótesis para determinar si existe multicolinealidad o no, por otra parte, se mide por el grado de correlación con el índice o factor de inflación de varianza (VIF) que sigue la siguiente fórmula:

$$\text{VIF}(x_i) = \frac{1}{1 - R_i^2} \tag{5.20}$$

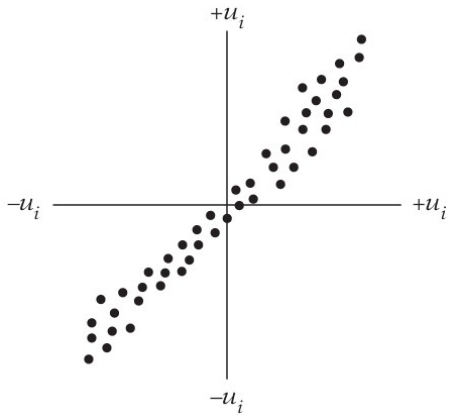


FIGURA 5.6: Ejemplo de residuos correlacionados.
Fuente: (Gujarati and Porter, 2010).

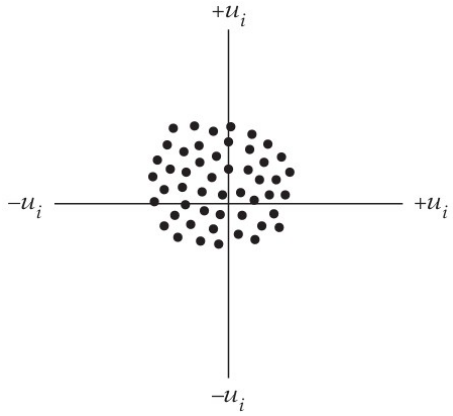


FIGURA 5.7: Ejemplo de residuos no autocorrelacionados.
Fuente: (Gujarati and Porter, 2010).

Donde R_i^2 es el coeficiente de correlación de un modelo generado tomando como variable respuesta la variable x_i. Un VIF aproximado a 1 significa correlación baja o nula, un VIF mayor a 5 significa un grado de correlación moderado o medio y un VIF mayor a 10 significa un grado de correlación severa o alta.

5.4.6. Prueba de Significancia General de una Regresión Múltiple.

Al igual que con los supuestos, el modelo obtenido también tiene una prueba de hipótesis para confirmar que el modelo se ajusta a los datos. La prueba en mención sigue la estructura:

$$\begin{aligned} H_o &: \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_a &: (\beta_0, \beta_1, \beta_2, \ldots, \beta_k) \neq 0 \end{aligned} \tag{5.21}$$

Dicha prueba sigue una distribución F cuyo valor de referencia se calcula así:

$$F = \frac{\text{SCE}/\text{gl}_{\text{SCE}}}{\text{SCR}/\text{gl}_{\text{SCR}}} \tag{5.22}$$

Donde SCE es la suma de cuadrados de la variable explicada, SCR es la suma de cuadrados de los residuos, gl_{SCE} son los grados de libertad de SCE dados por $k-1$, o en otras palabras, el número de estimadores y gl_{SCR} los grados de libertad de SCR dados por $n-k$, o en otras palabras, el número de observaciones menos el número de estimadores. Este valor se contrasta con el valor crítico $F_\alpha(k-1, n-k)$. En el caso en el que $F > F_\alpha$ se rechaza la hipótesis nula y se decide que el modelo se ajusta a los datos.

5.5. Regresión Lineal Múltiple por GLSAR.

El modelo GLSAR (*Generalized Least Squares with AutoRegressive errors*), o mínimos cuadrados generalizados con errores autorregresivos, es un tipo de regresión lineal principalmente utilizado en datos temporales, cuyos residuos no permanecen estáticos a lo largo de la muestra, sino que fluctúan con una determinada dinámica a medida que cambia la escala temporal, o en otras palabras, están autocorrelacionados. Este problema provoca que los coeficientes ya no sean insesgados y los vuelve ineficientes, puesto que no se está considerando toda la información disponible de manera apropiada. Por consiguiente, los errores estándar y las pruebas de significancia no son confiables (Buse, 1973).

5.5.1. Estructura del modelo.

Se inicia con la estructura de la regresión lineal simple para explicar la ecuación general del modelo GLSAR:

$$y_t = \beta_0 + \beta_1 x_t + e_t \tag{5.23}$$

Pero los residuos siguen una distribución autorregresiva de orden 1 o AR(1):

$$e_t = \rho e_{t-1} + u_t \tag{5.24}$$

Donde u_t es ruido blanco, ρ es un parámetro de autocorrelación para cada error . La ecuación (5.24) infiere que los errores están correlacionados a lo largo del tiempo. Por lo tanto, al reemplazar (5.24) en (5.23), y realizando una serie de procedimientos algebraicos, se obtiene que:

$$y_t - \rho y_{t-1} = \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + u_t \tag{5.25}$$

De acuerdo con la ecuación (5.25), los coeficientes de la regresión lineal se estiman con base en el resultado del error inmediatamente anterior. De esta manera, se corrigen los problemas de autocorrelación temporal en la muestra.

5.5.2. Matriz de Covarianza HAC.

Los errores estándar de los estimadores, en presencia de autocorrelación o heterocedasticidad, pueden presentar resultados erróneos, lo que se traduce en interpretaciones fallidas de la significancia estadística de los coeficientes. Por tal motivo, es necesario el uso de matrices de covarianza con errores robustos; de esta manera, se mejora la precisión de los residuos, lo que resulta en un análisis óptimo de los intervalos de confianza y pruebas de hipótesis en datos con componentes temporales,

residuos autocorrelacionados o sin varianza constante. Una de las matrices de covarianza robusta más utilizadas es la HAC (*Heteroscedasticity-Autocorrelation Consistent*, por sus siglas en inglés) la cual resume la varianza y la covarianza de una muestra de datos en múltiples subconjuntos (Andrews, 1991).

5.6. Modelo MARS.

El modelo de regresión con splines adaptativos multivariada es un algoritmo desarrollado por (Friedman, 1991) en el cual se crean modelos lineales por partes para las variables explicativas. Es una técnica no paramétrica y constituye, en esencia, una extensión de la regresión lineal sobre relaciones no lineales entre variables (Morales-Oñate, 2025). Algunas de las ventajas más representativas de este modelo son: la similitud, en cuanto a la forma, de las regresiones lineales tradicionales mientras se modelan relaciones e interacciones no lineales; la identificación de patrones inmersos en los datos; el modelado de relaciones parciales; la aplicación en variables tanto cualitativas como cuantitativas, y la transformación y selección automática de variables.

5.6.1. Comparación con las Regresiones Lineales y No Lineales Convencionales.

El principio básico de los modelos lineales parte del supuesto, como su nombre lo indica, de las relaciones lineales entre las variables; sin embargo, este enfoque limita la predicción de fenómenos con interacciones de otro tipo. Por lo tanto, se hace necesario extender los modelos para usar funciones polinómicas de la siguiente forma:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + e_i \tag{5.26}$$

De igual forma, para el correcto modelado del comportamiento de las variables, se puede hacer uso de funciones escalonadas, las cuales activan o desactivan los parámetros del modelo y siguen la siguiente forma:

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \beta_3 C_3(x_i) + \cdots + \beta_d C_d(x_i) + e_i \tag{5.27}$$

Donde $C_1(x_i)x_i c_x \leq c_2 C_2(x_i)x_i c_2 \leq x_i \leq c_3 \ldots C_d(x_i)x_i c_{d-1} \leq x_i < c_d$

El problema del desfase de los diferentes tipos de regresión lineal se ilustra en la figura (5.8). La regresión polinómica de grado 2 parcialmente resuelve el problema de desfase; no obstante, sigue presentando problemas al identificar los datos extremos. Por otra parte, la regresión polinómica de grado 3 sí permite representar el comportamiento de los datos de manera correcta. Asimismo, de manera gráfica, la regresión escalonada parcialmente modela el comportamiento de los datos; no obstante, el problema más relevante de este enfoque es que el analista debe, manualmente, identificar las funciones y los momentos en los que los datos cambian. Para conjuntos de datos muy grandes o relaciones entre variables que cambian significativamente, este proceso se vuelve extremadamente tedioso e innecesario.

5.6.2. Estructura Básica del Modelo.

El modelo MARS se basa en la identificación de relaciones no lineales de los datos a partir de la evaluación de puntos de corte (también llamados nudos) tal como lo hacen las funciones escalonadas. El proceso consiste en evaluar cada punto de datos para cada predictor como un nudo, y a partir de esa evaluación se ajusta una regresión lineal con las características candidatas. A manera de ejemplo, considérese un conjunto de datos no lineal en donde hay una variable Y que está en función de una variable X, es decir, $Y = f(X)$. El procedimiento MARS ejecuta un primer escaneo para identificar el punto en el que la relación de ambos conjuntos de datos alcanza el menor error, calculado con la sumatoria de los errores cuadráticos residuales ($\sum(y_i - f(x_i))^2$). El resultado de esa operación se conoce como función bisagra (F_b) y se representa así:

$$F_b = h(x - a) \tag{5.28}$$

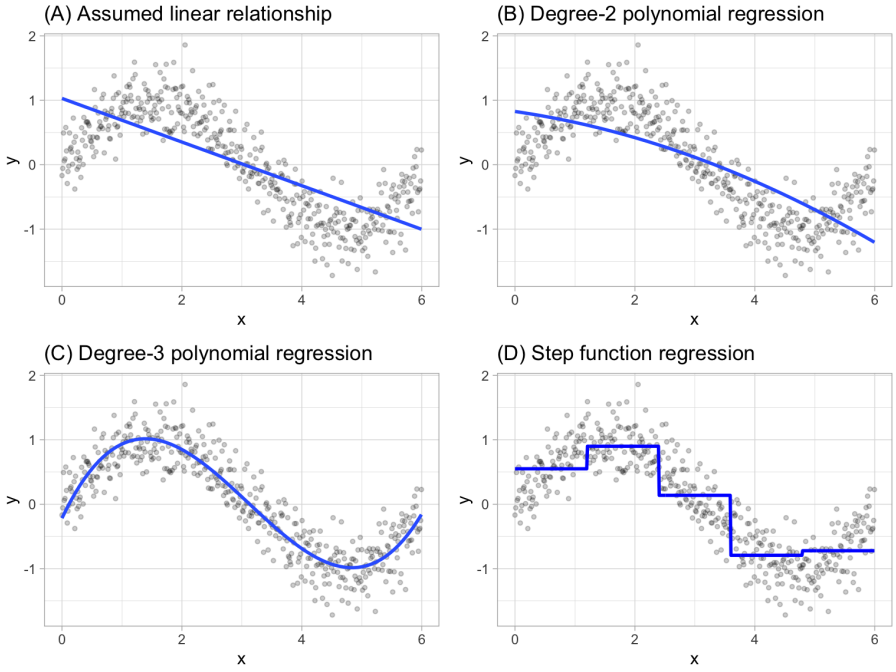


FIGURA 5.8: Ejemplos del comportamiento de la línea de regresión en datos con distribución polinómica de grado 3.
Fuente: (Boehmke and Greenwell, 2020).

Donde a es el valor en el que se produce el punto de corte. En este orden de ideas, se obtienen dos funciones o modelos lineales para un solo nudo:

$$y = \begin{cases} \beta_0 + \beta_1(a - x), & x < a \\ \beta_0 + \beta_1(x - a), & x > a \end{cases} \tag{5.29}$$

Nótese que, en el momento en que $x = a$ el parámetro β_i se cancela y no influye en el modelo. Siguiendo con el desarrollo del proceso, véase que la figura (5.9) ilustra gráficamente el modelo de la ecuación (5.29) con un a aproximado de $1,184$. Para un conjunto de datos con dos nudos, se obtendrían 3 modelos lineales representados de la siguiente manera:

$$y = \begin{cases} \beta_0 + \beta_1(a - x), & x < a \\ \beta_0 + \beta_1(x - a), & a < x < b, \\ \beta_0 + \beta_1(b - x), & x > b \end{cases} \tag{5.30}$$

Teniendo en cuenta los valores de la figura (5.9), los valores de a y b serían $1,184$ y $4,898$, respectivamente. Por lo que las ecuaciones (5.29) y (5.30) se expresarían de la siguiente forma:

$$y = \begin{cases} \beta_0 + \beta_1(1{,}184 - x), & x < 1{,}184 \\ \beta_0 + \beta_1(x - 1{,}184), & x > 1{,}184 \end{cases} \tag{5.31}$$

$$y = \begin{cases} \beta_0 + \beta_1(1{,}184 - x), & x < 1{,}184 \\ \beta_0 + \beta_1(x - 1{,}184), & 1{,}184 < x < 4{,}898, \\ \beta_0 + \beta_1(4{,}898 - x), & x > 4{,}898 \end{cases} \tag{5.32}$$

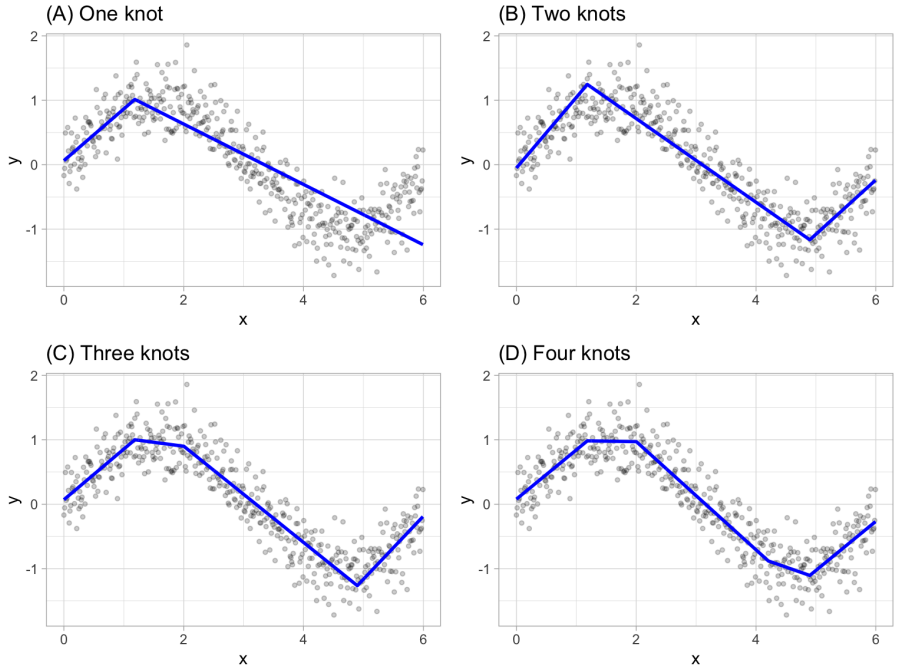


FIGURA 5.9: Ejemplos de conjuntos de datos con varios nudos.
Fuente: (Boehmke and Greenwell, 2020).

Por ende, el procedimiento continúa para todos los nudos que se encuentran, produciendo una ecuación final predictoria potencialmente no lineal de la forma:

$$y = f(x_i) = \beta_0 + \sum_{n=1}^{i} \beta_i B(x_i) \tag{5.33}$$

Donde y es la variable respuesta y β_i son los parámetros del modelo para las respectivas variables x_i que van desde $n = 1$ hasta i. El valor de β_0 representa el intercepto y $B(x_i)$ son las funciones bisagra de cada variable x_i.

Las funciones bisagra no necesariamente representan todas las relaciones o funciones lineales de los datos, debido a que pueden representar interacciones débiles que no aportan de manera significativa al modelo (Vanegas and Vásquez, 2017). Por tal motivo, dichas funciones pueden expresarse como $B(x_i) = \max(0, x_i - a)$ o $B(x_i) = \max(0, a - x_i)$. Los puntos de partición del espacio y los parámetros de los modelos se obtienen a partir de los datos analizados. El número de funciones base resultantes indica la complejidad del modelo

5.6.3. Método de Selección de Funciones.

El modelo MARS, como ya se mencionó anteriormente, se ejecuta en dos fases fundamentales:

- Fase o paso hacia adelante (*Forward Phase*): En esta fase, las funciones bisagra se agregan iterativamente, dejando aquellas que mejoran el ajuste del modelo.
- Fase o paso hacia atrás (*Backward Phase*): Aquellas funciones que son estadísticamente menos significativas para el modelo se eliminan para evitar el sobreajuste.

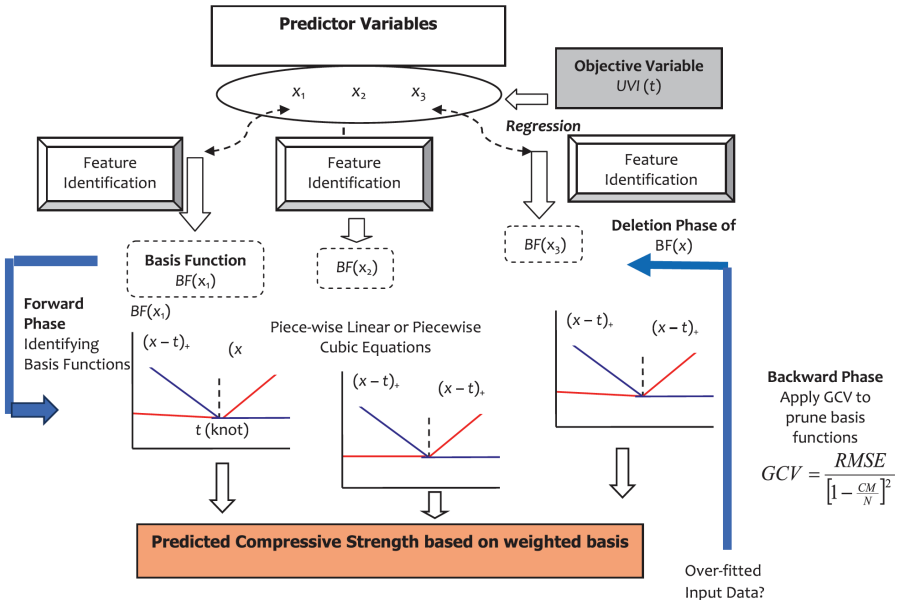


FIGURA 5.10: Arquitectura del modelo MARS.
Fuente: (Morales-Oñate, 2025).

El criterio de eliminación o adición de la regresión está basado en la metodología de validación cruzada generalizada, como lo indica la figura (5.10). Esta técnica utiliza una medida denominada GCV por las siglas en inglés de la metodología y se representa de la siguiente manera:

$$GCV = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\left(1 - \frac{C(m)}{n}\right)^2} \tag{5.34}$$

Donde y_i es el valor real de la variable respuesta, $\hat{y}_i$ es el valor estimado del modelo, n el número total de datos u observaciones y $C(m)$ es una medida de complejidad del modelo m, comúnmente igual al número de funciones bisagra más uno (por el intercepto). El numerador ya se había mencionado anteriormente para la determinación de las funciones bisagra y se le denomina el error cuadrático residual (RSS por sus siglas en inglés); el denominador penaliza la complejidad del modelo, es decir, cuando $C(m)$ se acerca a n, el denominador tiende a cero y el GCV tiende a infinito. Por lo tanto, el modelo con el menor GCV se considera el más balanceado, considerando una baja complejidad y un ajuste óptimo.

6

Diseño Metodológico.

En la presente sección se exponen las actividades requeridas para dar cumplimiento a los objetivos propuestos, teniendo en cuenta los siguientes tipos de metodologías de investigación:

- Método cuantitativo: se hace uso de datos numéricos y técnicas estadísticas para la construcción de relaciones entre variables con el fin de modelar su comportamiento en un ambiente determinado.
- Método correlacional: se estudia la relación entre variables mediante la obtención de ecuaciones o modelos sin determinar la causa directa.
- Método de modelado o simulación: se propone la construcción de un modelo que sintetice las relaciones entre las variables.

6.1. Zona de Estudio.

El lago George es un cuerpo de agua perteneciente a la cuenca del río San Juan Bautista, ubicado entre los condados de Volusia y Putnam al noreste del estado de Florida, Estados Unidos. Este lago se encuentra a 210 kilómetros del yacimiento del río y se caracteriza por ser el segundo cuerpo lacustre más grande en términos de extensión de superficie (por detrás del lago Okeechobee), con un total de 18.934 hectáreas y una profundidad promedio de tres metros (Piehler et al., 2009). El flujo dominante del lago proviene de la cuenca alta del río San Juan Bautista, con un 78,9 %. El porcentaje restante se divide entre escorrentía de cuencas hídricas locales, flujo manantial artesiano y precipitación directa.

6.1.1. Características Físicas.

La extensión total del lago es de 46.000 acres, lo que se representa en aproximadamente 186 kilómetros cuadrados. Cuenta con un largo máximo de 18 kilómetros y un ancho de 10 kilómetros. La profundidad media va de nueve a diez pies, o 2,7 a 3 metros. Debido a las características topográficas de esta zona específica del estado, el agua fluye hacia el norte, por lo que los cuerpos de agua que proveen al lago entran por el sur.

6.1.2. Topografía y Vida Acuática.

La topografía del lago George, en términos generales, es bastante consistente; a lo largo y ancho de todo el cuerpo, las variaciones en la profundidad son mínimas, rondando los tres metros en promedio. Adicionalmente, cuenta con una gran concentración salina, gracias a las fuentes de agua que lo proveen; dentro de las cuales se encuentran el lago Harney y los arroyos *Juniper*, *Glen* y *Salt*, lo que

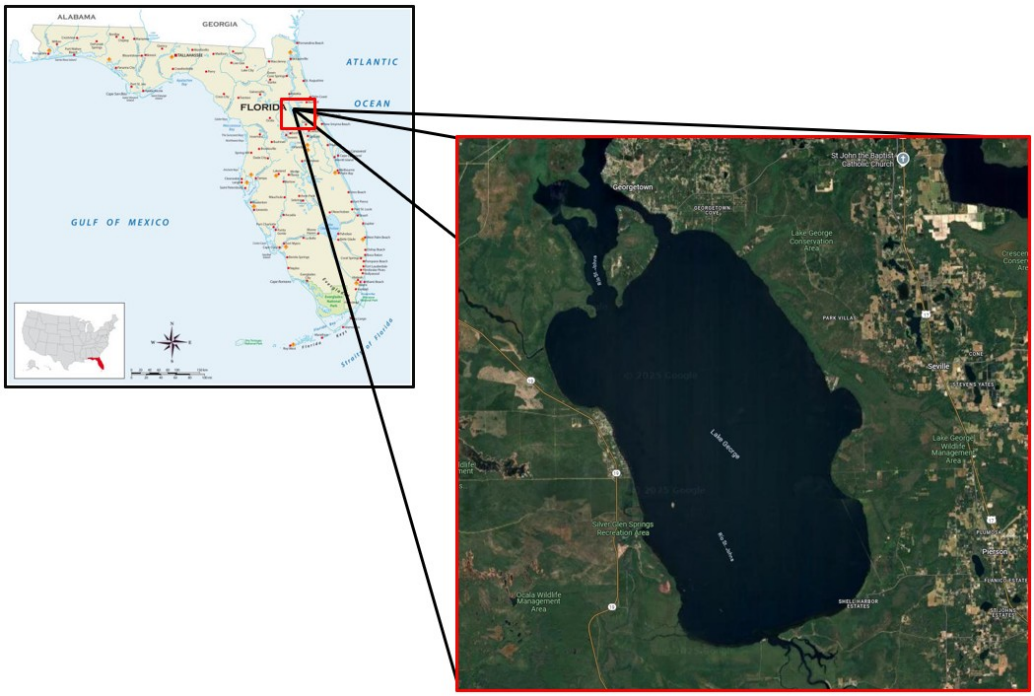


FIGURA 6.1: Localización del lago George.
Fuente: (Google Maps, 2026).

permite que numerosas especies de plantas y peces habiten y se reproduzcan en sus aguas. Con respecto a especies ictiológicas, las más representativas son la lubina negra y rayada, el pez sol de orejas rojas, la crappa negra y el pez gato marrón. Además, en la ribera se puede encontrar una especie local de cangrejo azul cuya actividad pesquera representa una parte importante de la economía regional (Porter, 2007).

Uso e Importancia Económica.

Diversas dinámicas se desarrollan actualmente en el lago, que generan réditos económicos, en su mayoría, de carácter privado. Dentro de las ofertas de actividades se encuentra el paseo en bote de vapor por la *Lake George Steamboat Company*, la cual cuenta con una gran variedad de vehículos acuáticos, incluyendo lanchas y cruceros. Debido a la diversa cantidad de especies piscícolas, la pesca deportiva es una actividad bastante representativa. En épocas secas, las atracciones con inflables son un atractivo turístico, incluyendo caminatas ecológicas, deportes acuáticos y ofertas de zonas de camping y navegación libre. En resumen, el lago tiene un impacto económico importante para los condados cercanos, tanto que el estado comenzó a dar licencias de construcción para desarrollos inmobiliarios residenciales en zonas delimitadas alrededor del cuerpo (Cumming, 2006).

Teniendo en cuenta el diagrama de flujo propuesto en la figura (6.2), las actividades que se realizaron para completar la investigación se dividieron en cuatro fases diferentes:

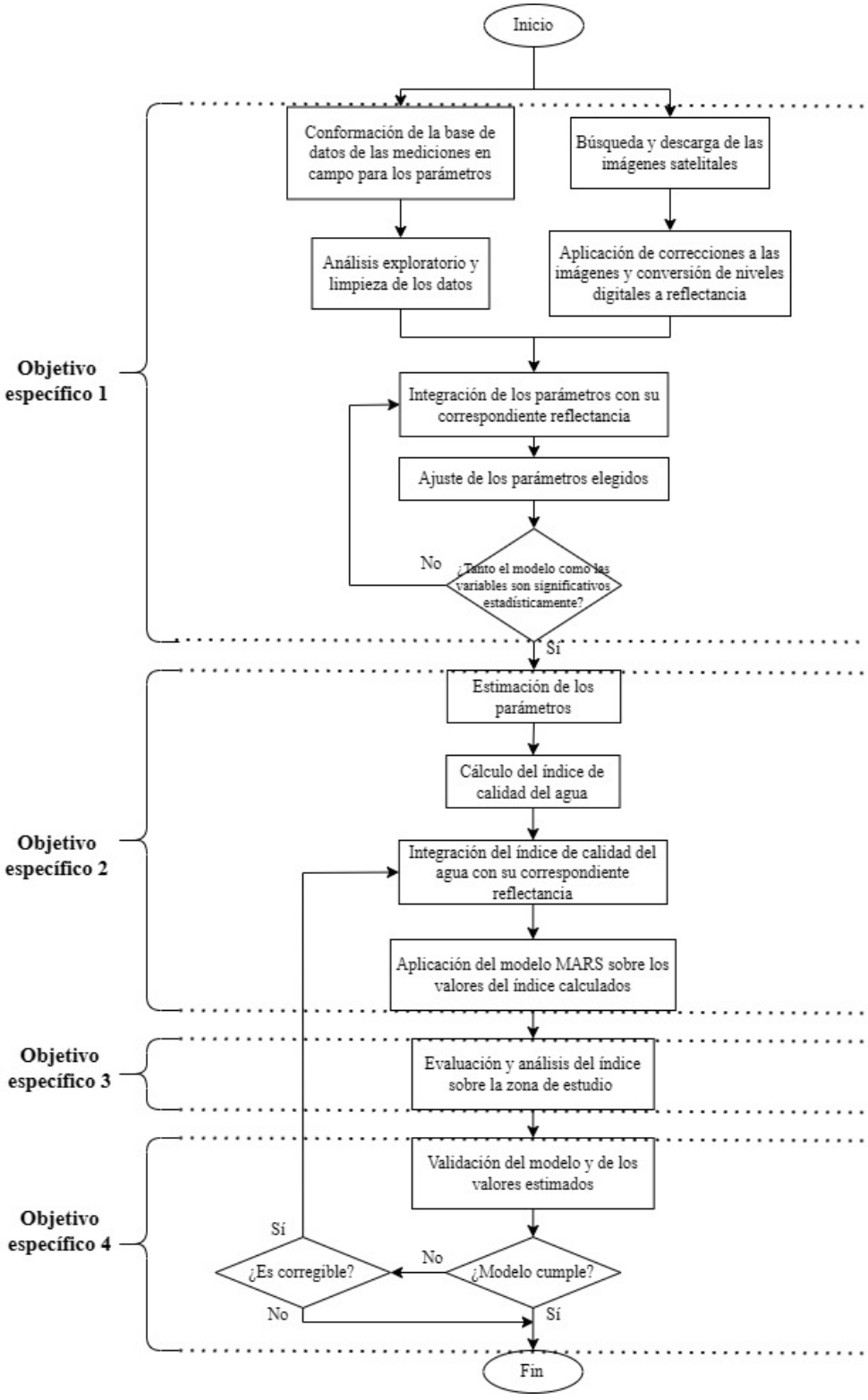


FIGURA 6.2: Metodología propuesta para la evaluación del índice de calidad del agua mediante el uso de imágenes Sentinel-2 y *machine learning*.
Fuente: Elaboración propia.

6.2. Fase 1: Preprocesamiento de los datos y obtención de las ecuaciones de los parámetros.

Este paso constó de dos procesos que se realizaron en paralelo. El primero se enfocó en la adquisición de las mediciones de cada uno de los 15 parámetros directamente del Atlas de Agua del Condado de Lake, al que se accedió en la fecha de realización del proyecto a través del siguiente enlace: https://lake.wateratlas.usf.edu/. Los 15 parámetros que inicialmente se procesaron fueron: arsénico, sulfato, cloruro, clorofila-a, oxígeno disuelto, dureza, nitrógeno, pH, fósforo, potasio, profundidad del disco Secchi, sodio, temperatura, sólidos suspendidos totales y turbidez. El segundo proceso se enfocó en la adquisición de las imágenes satelitales, las cuales se descargaron del geoportal de Copérnicus a través del siguiente enlace: https://dataspace.copernicus.eu/explore-data.

A las mediciones de los parámetros se les realizó un análisis exploratorio orientado a la descripción de su estructura, distribución y posibles patrones de comportamiento, mediante el lenguaje Python con la librería *pandas*, ideal para analizar datos numéricos. Algunos de los métodos utilizados de dicha librería fueron *mean()*, *median()*, *std()*, *min()* y *max()*. Adicionalmente, se uso la librería *scipy.stats* para calcular los coeficientes de asimetría y curtosis con los métodos *skew()* y *kurtosis()*, respectivamente. En esta parte del proceso también se obtuvieron los histogramas y diagramas *boxplot* por medio de la librería *seaborn*, la cual cuenta con los métodos *histplot()* y *boxplot()*, que generan automáticamente las gráficas requeridas.

Asimismo, se llevó a cabo un proceso de limpieza mediante la metodología del rango intercuartílico, asegurando la integridad de los datos para su posterior procesamiento. Adicionalmente, con el software SNAP, se procesaron las bandas de las imágenes satelitales mediante las herramientas de geoprocesamiento ráster de tipo geométrico. En este paso se recortaron las imágenes para enfocarse solamente en el área de estudio y se aplicaron correcciones de tipo geométrico, atmosférico y radiométrico, de tal forma que los recortes no extendieran ningún tipo de error a lo largo de la cadena de procesos. Finalmente, se aplicó un proceso de remuestreo a las bandas con una resolución espacial diferente a 10 metros, de tal forma que la geometría de todos los píxeles fuera igual a lo largo de toda la muestra.

Luego del preprocesamiento tanto de las mediciones como de las imágenes satelitales, se procedió a relacionar los valores de reflectancia de cada banda con los valores de concentración de cada parámetro de la siguiente forma: se desplegaron las imágenes mediante la librería *rasterio* y el método *open()*; luego, se codificaron las coordenadas de las estaciones de tal forma que se convirtieran en un objeto tipo punto con el método *Point()* de la librería *geopandas* de Python; finalmente, se relacionaron las mediciones de los parámetros con sus correspondientes valores espectrales con base en la fecha de obtención de cada insumo. Este proceso fue corroborado manualmente por medio del software *ArcGIS Pro*, añadiendo aleatoriamente una banda y validando el valor extraído versus el valor leído en pantalla.

Posteriormente, se realizaron múltiples regresiones para determinar cuál era la combinación de bandas que mejor se ajustaba a las mediciones de campo. Dichas combinaciones se realizaron con la ayuda del método *combinations()* de la librería *itertools*, de tal forma que se obtuvieron todos los conjuntos de bandas posibles para cada modelo. Esta tarea se complementó con la librería *statsmodels* de Python, desarrollada para la aplicación de regresiones. El cálculo de las medidas básicas de validación estadística para modelos de este tipo y la obtención de las pruebas de normalidad, heterocedasticidad y autocorrelación de los residuos se realizaron con las librerías *statsmodels* y *scipy.stats* utilizando los métodos *shapiro()*, *kstest()*, *jarque_bera()*, *het_breuschpagan()*, *het_white()* y *het_goldfeldquandt()*.

6.3. Fase 2: Estimación de los parámetros y modelado del índice de calidad del agua.

Una vez que se completaron y validaron las ecuaciones para los parámetros correspondientes, se procedió a aplicarlas en zonas adyacentes a las estaciones de monitoreo en las cuales todos los modelos eran evaluables para poder determinar los valores de cada indicador. Esta actividad requirió la aplicación de diferentes métodos relacionados con las librerías *statsmodels*, *rasterio*, *geopandas*, *pandas* y *numpy*. En consecuencia, resultó en la conformación de una base de datos de parámetros estimados con la cual se calculó el índice de calidad del agua bajo los lineamientos del CCME (Neary et al., 2001). El proceso anterior permitió la obtención de un valor del índice para múltiples zonas del lago George. A continuación, se relacionaron dichos valores con los valores de reflectancia correspondientes a cada una de las bandas de las imágenes seleccionadas. Estos últimos conjuntos de datos consolidados sirvieron de insumo para implementar el modelo MARS; este ajuste se realizó con la librería *py-earth* y los métodos *fit()* y *predict()*, los cuales están diseñados específicamente para el ajuste flexible de modelos complejos no lineales.

6.4. Fase 3: Evaluación del modelo.

Con el ajuste final del modelo propuesto, fue posible aplicarlo en diferentes zonas del Lago para estimar valores del índice con ayuda de la misma librería *py-earth*, utilizando solamente los valores de las bandas de las imágenes satelitales dispuestas para su evaluación. Dado que el estudio se realizó en el año 2023, el modelo se evaluó en las mismas imágenes utilizadas en las actividades mencionadas anteriormente, con el fin de comparar los valores estimados con los calculados en la fase 2. En este sentido, también se analizó de manera práctica la evaluación del modelo a través de la creación de mapas temáticos del índice estimado utilizando el software *ArcGIS Pro*, interpretando los valores con respecto a la escala de calidad del CCME.

6.5. Fase 4: Validación del modelo.

La fase final consistió en la validez del modelo desde un punto de vista estadístico; téngase en cuenta que el ajuste se realizó con el 70 % de los datos; el restante 30 % se empleó en la fase de validación. En primer lugar, se analizaron las estadísticas básicas de ajuste del modelo, tales como el coeficiente de determinación R^2, el error medio cuadrático (RMSE), el error medio absoluto (MAE) y el error relativo. Posteriormente, el modelo se sometió a un proceso de validación cruzada, *train-test split* y *bootstrap* con los datos reservados para tal fin. Finalmente, se analizaron los residuos arrojados por el modelo mediante gráficas de distribución, gráficas comparativas entre los valores del índice estimados y los residuos correspondientes, además de la interpretación de cada variable y su importancia en el modelo. La serie de procesos anteriores se llevó a cabo con las librerías *sklearn*, *py-earth* y *matplotlib* de Python.

7

Descripción y Análisis Exploratorio de Datos.

Las variables seleccionadas para el desarrollo de la presente investigación corresponden a parámetros de calidad del agua que tienen una incidencia directa en el sabor, la claridad y la integridad físico-química del lago, cuya cantidad de registros era lo suficientemente robusta para realizar modelos de regresión lineal con resultados satisfactorios; por lo tanto, es posible que una porción significativa de minerales y variables no fuera tenida en cuenta, ya sea por la limitada cantidad de mediciones en la zona de estudio o por la ausencia de registros disponibles en las fuentes de datos abiertos consultadas.

7.1. Parámetros de Calidad del Agua.

Para el desarrollo inicial de los modelos de regresión lineal se tomó en cuenta un total de 15 parámetros cuyas mediciones se realizaron en el período comprendido entre el 1 de septiembre de 2022 y el 30 de abril de 2024. Dichas mediciones fueron descargadas del atlas de recursos hídricos del condado de Lake: https://lake.wateratlas.usf.edu/data-download/. Los datos fueron recolectados de dos estaciones identificadas con los números 44672 y 44673, ubicadas en la parte central del lago. Los parámetros seleccionados se listan a continuación:

- Arsénico (μg/l)
- Sulfato (mg/l)
- Clorofila (μg/l)
- Cloruro (mg/l)
- Oxígeno disuelto (DO) (mg/l)
- Dureza (mg/l)
- Fósforo (mg/l)
- Nitrógeno (mg/l)
- pH
- Potasio (mg/l)
- Profundidad del disco Secchi (SDD) (pies)

- Sodio (mg/l)
- Temperatura (°F)
- Sólidos suspendidos totales (TSS) (mg/l)
- Turbidez (NTU)

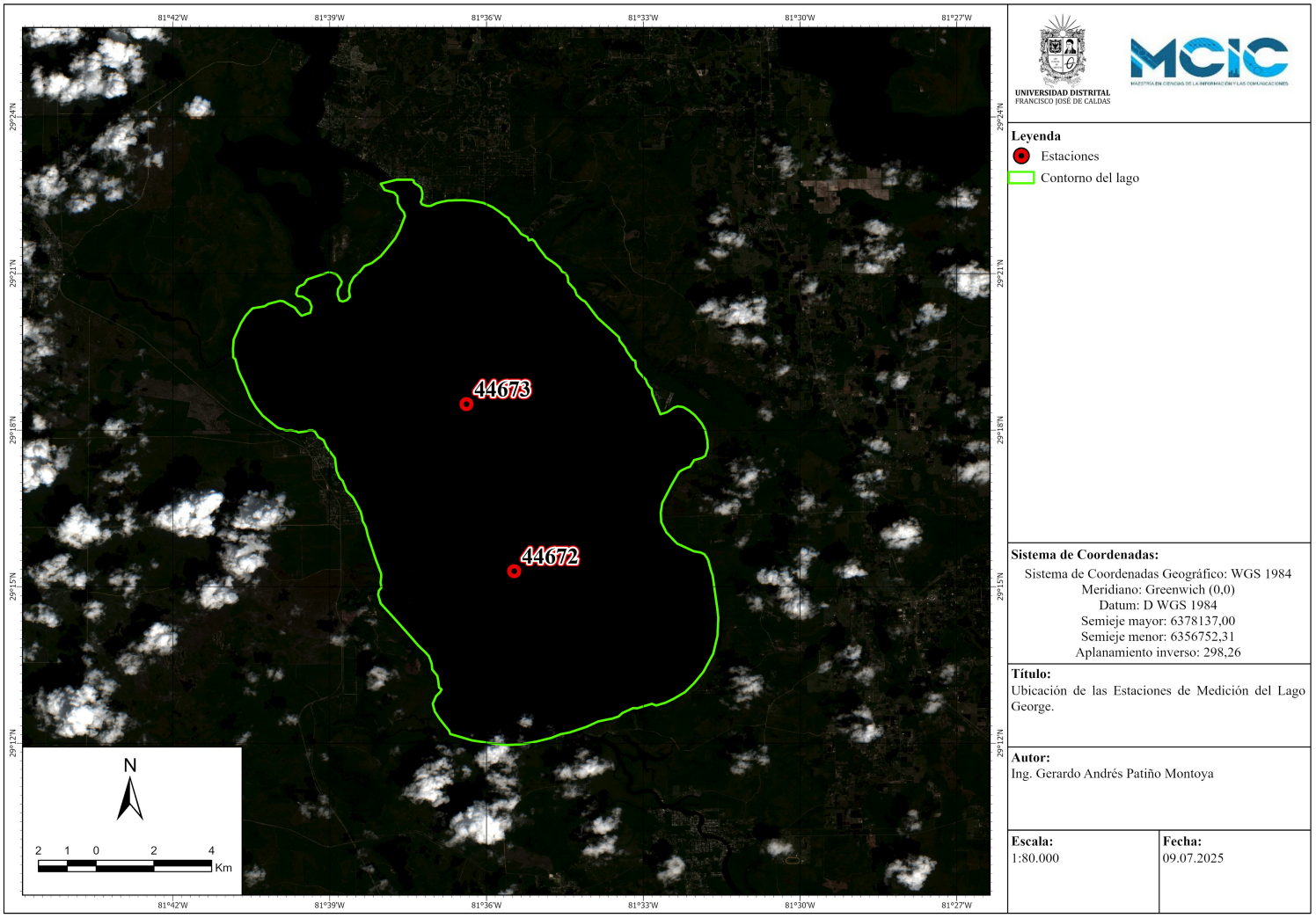


FIGURA 7.1: Ubicación de las estaciones en el Lago George.
Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023).

7.1.1. Estadísticas descriptivas.

En la tabla (7.1) se resumen las estadísticas descriptivas calculadas para cada uno de los parámetros elegidos. Analizando la media y la mediana de los valores, se puede considerar que los valores son consistentes con respecto al teorema del límite central, el cual indica si una variable tiene un comportamiento similar a la distribución normal. Si bien no es completamente necesario que los parámetros sigan dicho comportamiento, sí da una noción de la variabilidad y posible distribución de los datos para poder ser ajustados a un modelo de regresión lineal. Nótese que prácticamente la mitad de los parámetros presenta un sesgo hacia la izquierda, sustentado por un coeficiente de asimetría negativo. Además, el cloruro, el sodio y la dureza son los parámetros con mayor variabilidad; por otra parte, el fósforo, el arsénico y el nitrógeno son los minerales con menor variabilidad en la base de datos final.

En términos generales, las estadísticas descriptivas indican que la mayoría de los valores de los parámetros tienen un comportamiento normal con un sesgo ligeramente menor; no obstante, el sulfato, la clorofila-a, el cloruro, el sodio, la temperatura y la turbidez muestran rangos considerablemente amplios, lo que indica que las mediciones de estas variables están fuertemente influenciadas por un componente adicional, como lo puede ser el temporal, en el sentido en que la actividad biológica y eutrófica se incrementa en temporadas con temperaturas más altas; es decir, en el periodo de verano; a su vez, disminuyen en temporadas de invierno, con sensaciones térmicas más bajas. La curtosis calculada es la de exceso; por lo tanto, el valor de referencia es cero (no tres según la definición de curtosis clásica). En este orden de ideas, una curtosis cercana a cero significa comportamiento normal; una curtosis mayor a cero significa una distribución de valores más concentrada en el centro de los datos, o, una ausencia de valores extremos; por último, una curtosis menor a cero indica mayor presencia de valores extremos, lo que sugiere mayor probabilidad de valores atípicos y un grado mayor de variabilidad.

En consecuencia, el potasio, los sólidos suspendidos totales, el sulfato y el cloruro constituyen el grupo de parámetros con mayor tendencia a presentar concentraciones más bajas alrededor de la media, lo que influye en la obtención de valores considerablemente dispersos. Por otra parte, la mayoría de las variables mostraron un valor absoluto de curtosis en el rango de 0,8 a 1,0; por lo tanto, fue posible identificar una dispersión uniforme dentro de la propia dinámica de las mediciones. Finalmente, la clorofila-a y la SDD demostraron ser los indicadores con la mayor concentración de observaciones alrededor de la media, con una curtosis de 0,1725 y 0,1268, respectivamente.

TABLA 7.1: Estadísticas descriptivas de los parámetros medidos en el Lago George.

Parámetro	**Media**	**Mediana**	**Desv. Est.**	**Mínimo**	**Máximo**	**Asimetría**	**Curtosis**
Arsénico	0,6896	0,6881	0,1202	0,4786	0,9177	-0,1055	-0,8188
Sulfato	49,3163	48,0211	19,2330	16,0311	83,4388	-0,0240	-1,1674
Clorofila-a	24,5209	23,8027	12,8227	2,9904	58,8135	0,56	0,1725
Cloruro	200,0083	202,63	62,0528	77,34	288,97	-0,2925	-1,1706
DO	7,7764	7,8	0,9143	5,85	9,8	0,0481	-0.8265
Dureza	161,5748	174,045	40,0987	69,9801	208.8604	-0,5125	-0.8129
Fósforo	0,0554	0,0484	0,0368	-0,0033	0,1457	0,4674	-0.8334
Nitrógeno	1,3051	1,2779	0,2133	0,8813	1,6843	0,1796	-0.6004
pH	7,9662	7,925	0,5156	6,55	8,95	-0,0749	-0.7673
Potasio	4,5462	4,7881	0,9826	2,9083	5,9453	-0,2496	-1.4456
SDD	2,1254	1,9685	0,7694	0,6562	3,937	0,7087	0.1268

Parámetro	Media	Mediana	Desv. Est.	Mínimo	Máximo	Asimetría	Curtosis
Sodio	110,4181	111,6081	36,6243	40,5064	166,7091	-0,0629	-1.0354
Temperatura	75,5355	76,0046	8,5255	56,12	89,1338	-0,4457	-0.9711
TSS	11,2788	11,0	4,8509	1,6	27,0	0,7377	1.3097
Turbidez	5,6874	4,7123	2,8386	1,4306	14,0742	1,1541	1.0052

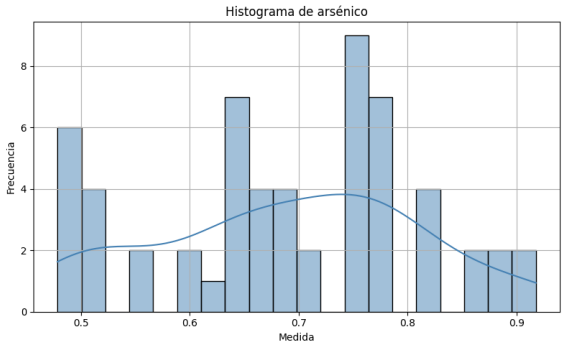
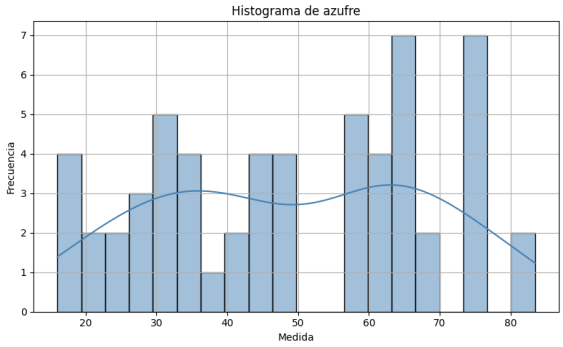
Fuente: Elaboración propia.

7.1.2. Histogramas.

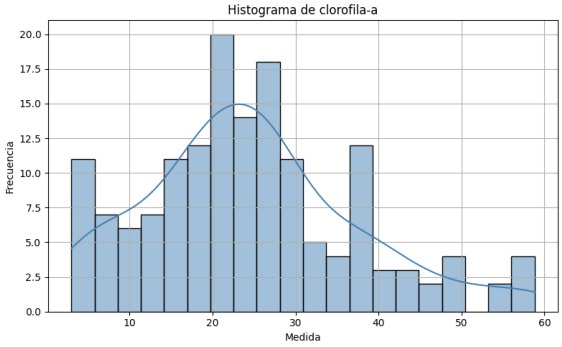
La tabla (7.2) muestra los histogramas de cada uno de los parámetros finales. En las diferentes gráficas se puede evidenciar que los valores medidos para los parámetros son bastante variados, teniendo comportamientos atribuibles a la distribución normal o simétricos, así como patrones bimodales y curvas sesgadas hacia la derecha o hacia la izquierda. El oxígeno disuelto y nitrógeno, por ejemplo, presentan un comportamiento simétrico con varias medidas concentradas en los valores centrales. Los parámetros que presentan un sesgo hacia la izquierda son: clorofila-a, fósforo, tss y turbidez. Por otra parte, la temperatura y la dureza muestran un sesgo hacia la derecha donde la mayoría de las medidas están concentradas en los valores más altos. Por último, hay patrones bimodales o con doble pico, como es el caso del sulfato, el cloruro, el pH y el potasio; en este tipo de histogramas se evidencia que los datos se concentran en dos valores distintos.

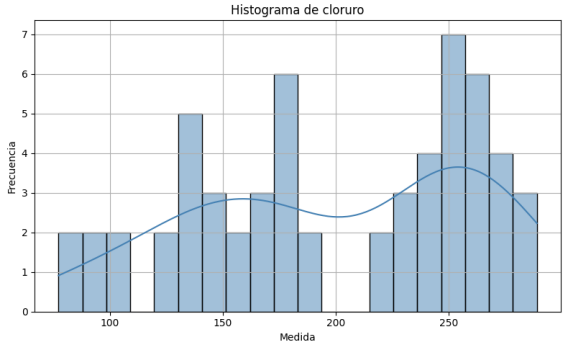
Adicionalmente, el sodio, la SDD y el arsénico cuentan con un comportamiento mayormente distribuido a lo largo del histograma, donde sus valores más frecuentes se distribuyen en diferentes medidas. Para entender un poco mejor la variabilidad y concentración de la mayoría de los valores de las medidas, véase la sección siguiente en donde se muestra el diagrama de caja y bigotes o *Box-Plot*, en el cual se podrán analizar aquellos valores que se pueden considerar como *outliers* o atípicos.

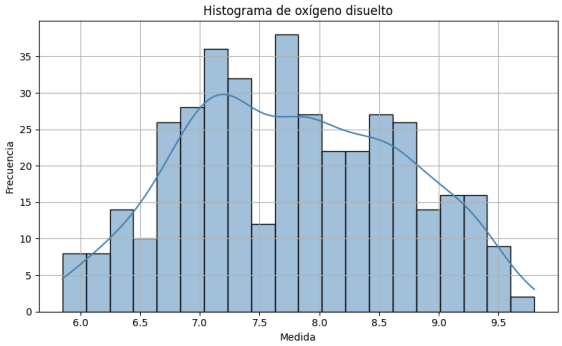
TABLA 7.2: Histogramas de los parámetros medidos en el Lago George.

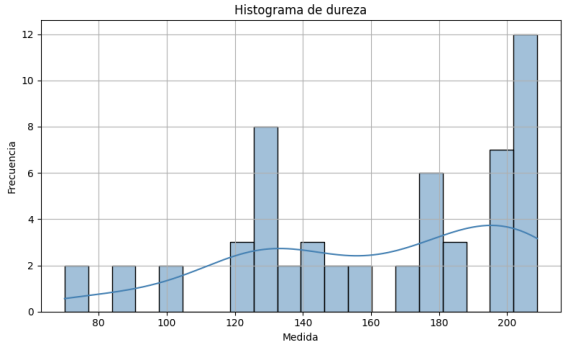
Histogramas
 
Continúa en la siguiente página

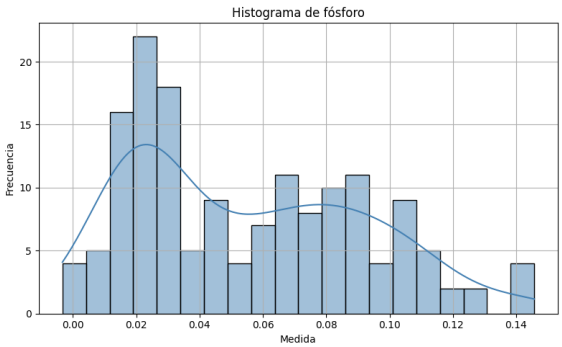
Histogramas

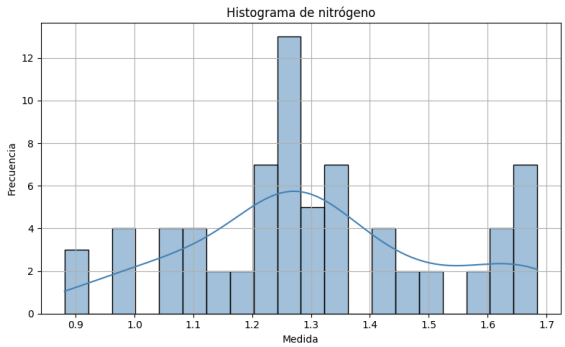


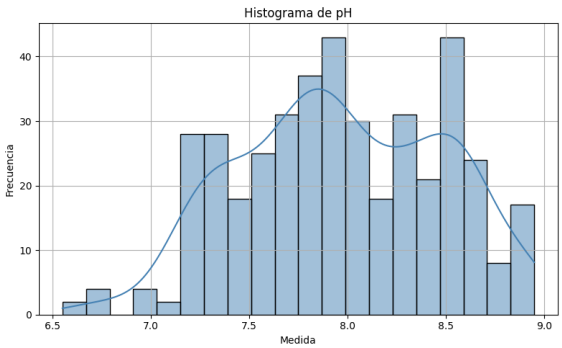


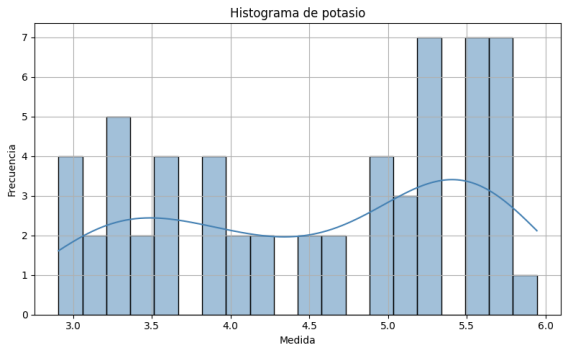






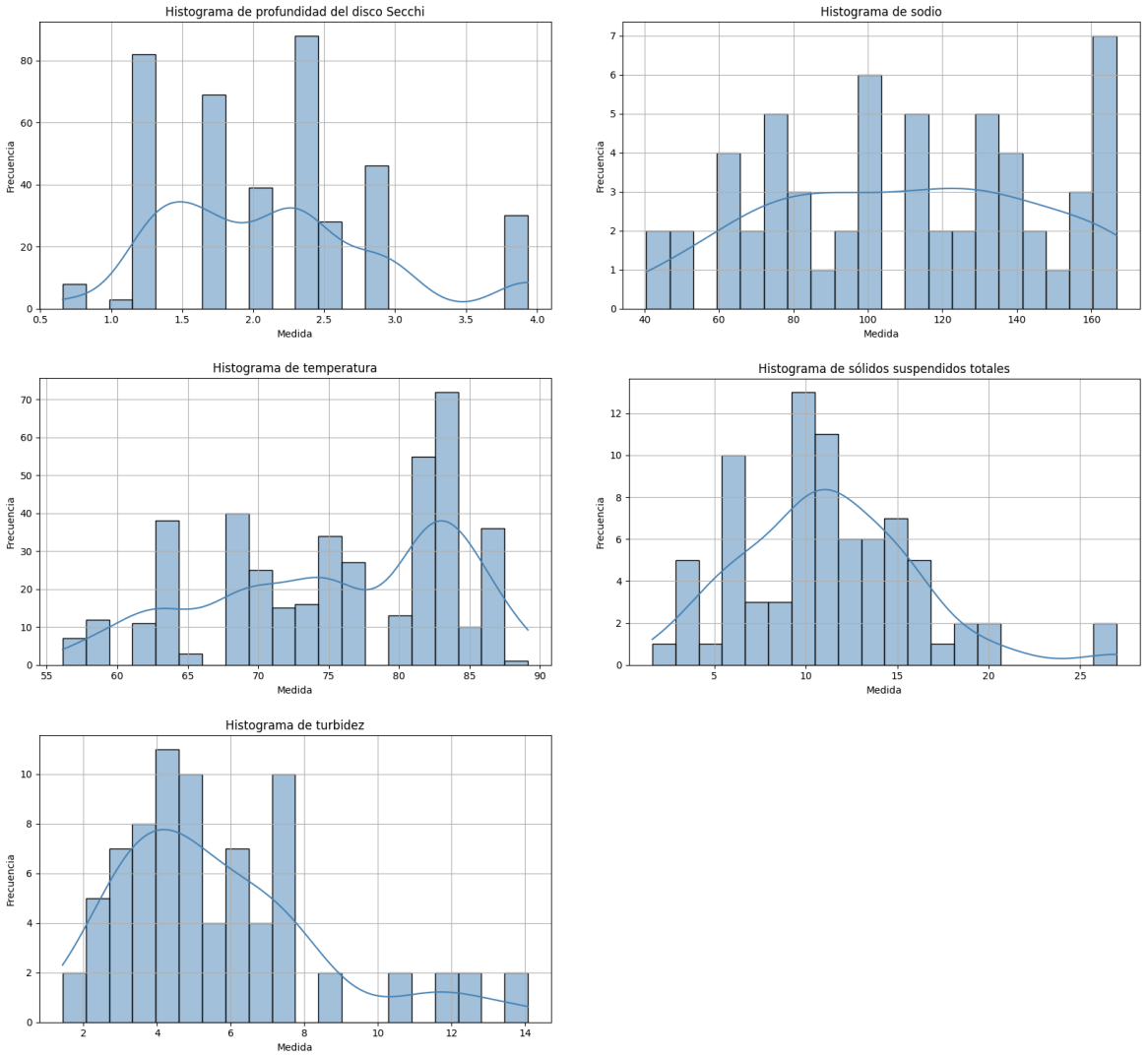






Continúa en la siguiente página

Histogramas



Fuente: Elaboración propia mediante la librería *seaborn* de Python.

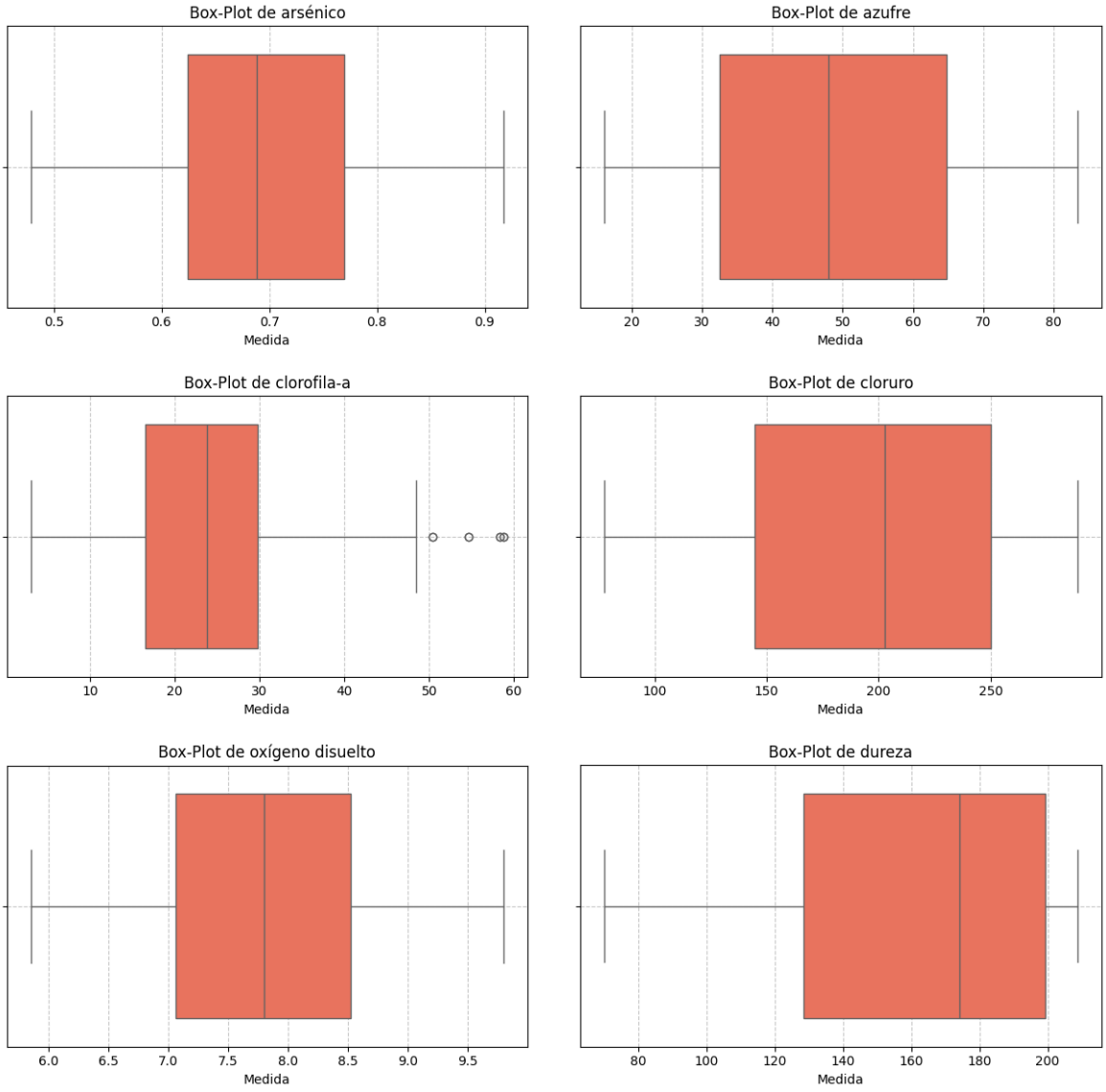
7.1.3. Diagrama *Box-Plot*.

El diagrama de caja y bigotes, o *Box-Plot*, permite analizar el grado de variabilidad y la concentración de un grupo de datos a través de la organización de estos en cuartiles (Williamson et al., 1989). Este diagrama es una alternativa al análisis descriptivo realizado con los histogramas, debido a que permite identificar patrones de sesgo, por ejemplo. Parámetros como la clorofila-a, la turbidez, los TSS y el fósforo evidencian que una gran parte de los datos se encuentra en los primeros 3 cuartiles. Los parámetros que cuentan con un comportamiento más equilibrado, es decir, con una variabilidad más distribuida a lo largo de los cuartiles, son: el arsénico, el sulfato, el DO y el sodio; estos diagramas muestran que tanto la caja como los bigotes tienen una longitud similar, por lo que el grado de variabilidad en estos parámetros es menor.

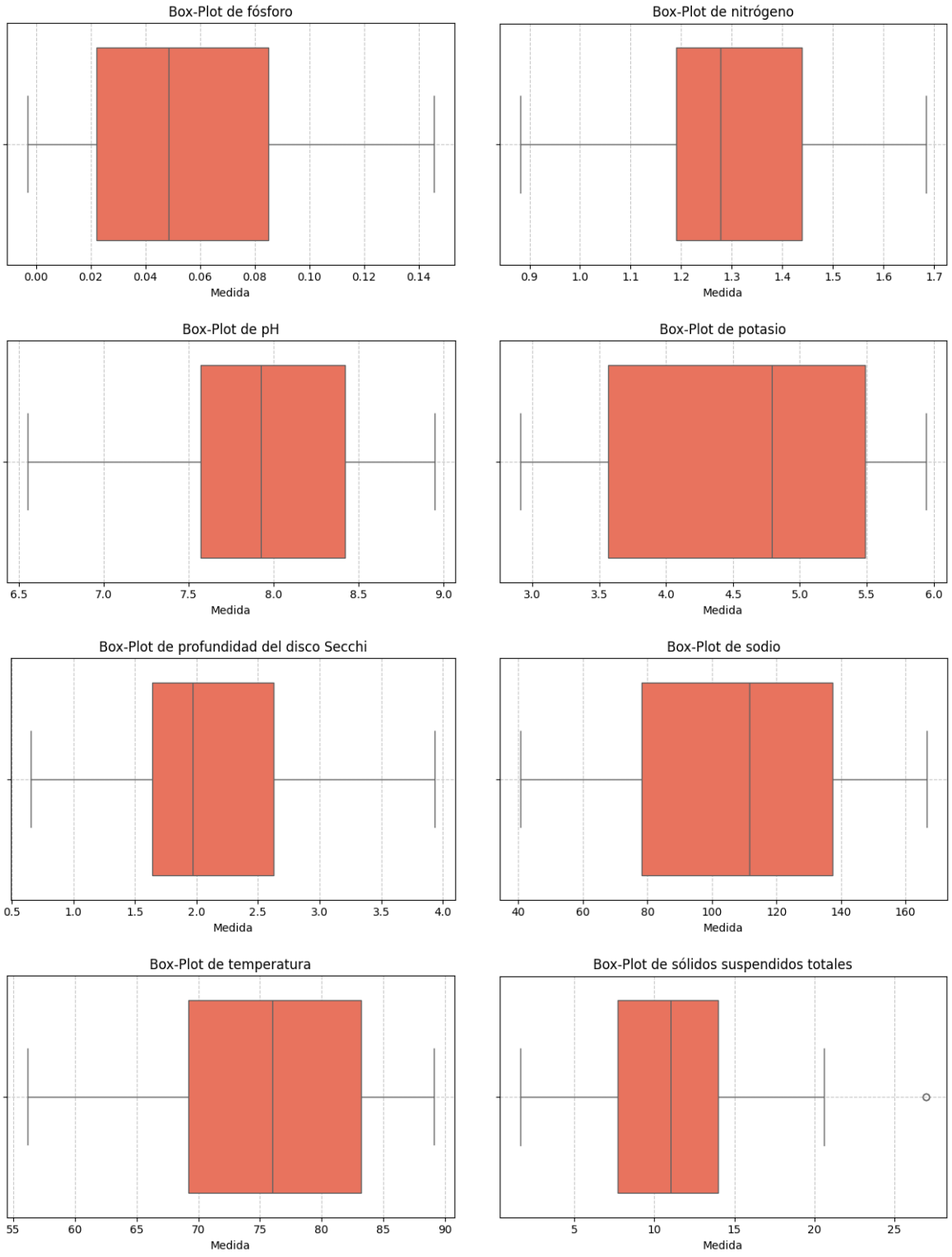
Por otra parte, hay parámetros que presentan una mayor concentración de datos, ya sea en la caja (cuartil 25 al 75) o en los bigotes, como es el caso del nitrógeno o la SDD, cuyos bigotes son más

alargados en comparación con su caja correspondiente, lo que indica que hay una mayor variabilidad en los extremos de las medidas. El potasio evidencia el caso contrario, en el cual el tamaño de la caja es en proporción más grande que la longitud de los bigotes, por lo que se puede concluir que la variabilidad de los datos está más concentrada en los valores medios que en los extremos. Por último, la clorofila-a, los TSS y la turbidez son las únicas variables que presentan valores atípicos obtenidos por medio del método del diagrama *Box-Plot*, siendo esta última la que tiene la mayor cantidad de *outliers* con cuatro.

TABLA 7.3: *Box-Plot* de los parámetros medidos en el Lago George.

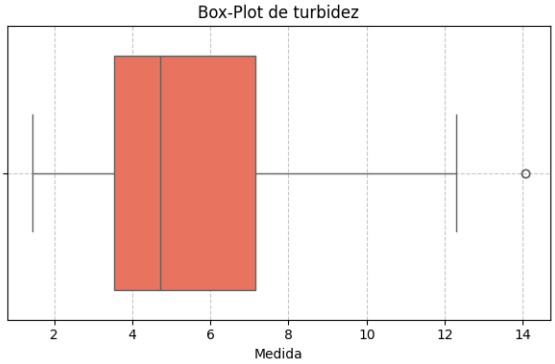
Box-Plot

Continúa en la siguiente página

Box-Plot



Continúa en la siguiente página

Box-Plot



Fuente: Elaboración propia mediante la librería *seaborn* de Python.

7.2. Imágenes Sentinel-2.

Se utilizaron en total 20 imágenes satelitales del programa Sentinel-2, descargadas directamente desde *Copérnicus*, que es un geoportal de la Agencia Espacial Europea (ESA, European Space Agency -por sus siglas en inglés-), y se encuentra disponible en el siguiente enlace: `https://browser.dataspace.copernicus.eu/` (visitado el 27 de agosto de 2025). Debido a que todas las imágenes se descargaron con el nivel más alto de procesamiento, no fue necesario pos-procesar las bandas obtenidas. El rango de tiempo elegido para el desarrollo del proyecto comprende datos medidos para cada parámetro desde septiembre de 2022 hasta abril de 2024. En consecuencia, se obtuvo una imagen satelital para cada mes dentro de dicho periodo. En la tabla (7.4) se resume el nombre del producto junto con la fecha de adquisición correspondiente.

TABLA 7.4: Producto y fecha de adquisición de las imágenes.

Producto	**Fecha de adquisición**
S2A_MSIL2A_20220920T155941	20-09-2022
S2B_MSIL2A_20221025T160339	25-10-2022
S2A_MSIL2A_20221129T160621	29-11-2022
S2A_MSIL2A_20221219T160701	19-12-2022
S2A_MSIL2A_20230108T160641	08-01-2023
S2A_MSIL2A_20230217T160311	17-02-2023
S2A_MSIL2A_20230309T160101	09-03-2023
S2A_MSIL2A_20230418T155821	18-04-2023
S2B_MSIL2A_20230513T155829	13-05-2023
S2B_MSIL2A_20230602T155819	02-06-2023
S2B_MSIL2A_20230712T155829	12-07-2023
S2B_MSIL2A_20230811T155829	11-08-2023
S2B_MSIL2A_20230910T155829	10-09-2023
S2B_MSIL2A_20231020T160209	20-10-2023
S2B_MSIL2A_20231109T160459	09-11-2023
S2B_MSIL2A_20231219T160659	19-12-2023
S2B_MSIL2A_20240128T160529	28-01-2024

Producto	Fecha de adquisición
S2B_MSIL2A_20240207T160429	07-02-2024
S2A_MSIL2A_20240323T155921	23-03-2024
S2B_MSIL2A_20240407T155819	07-04-2024

Fuente: Elaboración propia.

8

Desarrollo de los Modelos de Regresión Lineal.

La siguiente etapa del proceso comprendía la consolidación de una base de datos que relacionara las medidas tomadas por las estaciones en el lago y la subsecuente respuesta espectral de cada imagen. Así entonces, se agruparon las observaciones por mes y se sustrajeron los valores de reflectancia de cada banda, asignándoselos a su observación correspondiente según su fecha. Para tal fin, fue necesario procesar las bandas de tal forma que mantuvieran la misma resolución espacial. A este proceso se le conoce como remuestreo, tomando como capa base la banda 2 (Azul), que naturalmente tiene una resolución de 10 metros. En ese orden de ideas, las bandas 1, 5, 6, 7, 8A, 9, 11 y 12 pasaron de tener resoluciones entre 20 y 60 metros a 10 metros, como las bandas del espectro visible.

8.1. Limpieza y Filtrado de Variables Significativas.

Los valores de reflectancia de las imágenes utilizadas, como ya se mencionó, fueron sujetos a una fase de posprocesamiento correspondiente al nivel 2 para una imagen Sentinel, lo que quiere decir que el rango esperado para dichos valores está entre 0 y 1 al ser números escalados, siendo 0 los niveles más oscuros relacionados con superficies que absorben una gran cantidad de energía y, por consiguiente, no reflejan la luz, y 1 los niveles más claros relacionados con elementos de la superficie que reflejan la mayoría de la energía, absorbiendo una cantidad minúscula de ésta (Wang et al., 2016); no obstante, debido a que la superficie de análisis corresponde a valores muy pequeños, se hallaron valores negativos, lo que en la práctica no sería lógico; sin embargo, este fenómeno se ha documentado y analizado en múltiples estudios (ESA, 2021; Warren et al., 2019; Vanhellemont, 2019; Pereira-Sandoval et al., 2019), justificado por el hecho de que las correcciones atmosféricas y geofísicas de los productos de nivel 2 afectan de manera directa los valores extremos de los valores de reflectancia.

En ese orden de ideas, plantear un modelo de regresión lineal múltiple supondría el hallazgo de problemas de multicolinealidad ocasionados por la similitud de los valores de reflectancia. Por tal motivo, se implementó una metodología utilizada previamente por (Najafzadeh and Basirian, 2023), la cual consiste en la determinación de índices normalizados y ratios o cocientes con cada una de las bandas con la siguiente estructura:

$$ND_{ij} = \frac{B_i - B_j}{B_i + B_j} \tag{8.1}$$

$$R_{ij} = \frac{B_i}{B_j} \tag{8.2}$$

En donde ND_{ij} es el índice entre las bandas i y j, R_{ij} es el ratio o cociente entre las bandas i y j, B_i y B_j son los valores de reflectancia de las bandas i y j.

8.1.1. Limpieza de los Datos.

Previo a cualquier tipo de procesamiento, se realizó una rutina de limpieza de datos básica mediante la técnica del rango intercuartílico (Walfish, 2006), la cual se utiliza, por ejemplo, en la determinación de valores atípicos en las gráficas de caja y bigotes, también conocidas como diagramas *Box-Plot*. Esta técnica halla la diferencia entre el cuantil 75 y 25; luego, determina un rango multiplicando dicha diferencia por un factor, el cual, para el ejercicio, fue de 1,5; finalmente, todos aquellos valores que se encuentren fuera del rango obtenido se consideran atípicos y se eliminan del análisis.

$$IQR = Q_3 - Q_1 \tag{8.3}$$

$$L_i = Q_1 - F * IQR \tag{8.4}$$

$$L_s = Q_3 + F * IQR \tag{8.5}$$

Donde IQR es el rango intercuartílico, Q_3 es el valor del cuantil 75, Q_1 es el valor del cuantil 25, F es el factor de multiplicación del IQR, L_i y L_s son los límites inferior y superior, respectivamente.

La table (8.1) presenta el resumen de cada uno de los datos sustraídos en el análisis de datos atípicos, incluyendo el tamaño original, las observaciones resultantes, la cantidad de datos eliminados y el porcentaje de datos eliminados. Nótese que los parámetros con la mayor cantidad de mediciones (pH y temperatura) fueron aquellos a los que se les removió la menor cantidad de datos, con menos del 10 %. Además, los parámetros que presentaron una mayor cantidad de variabilidad, analizados desde el punto de vista de valores atípicos, fueron la clorofila-a y la turbidez, a los que se les removió un 33,3 % y 37,18 % de los registros, respectivamente. Para los parámetros relacionados con nutrientes como el arsénico, sulfato, cloruro, potasio y sodio, se evidenció que no solo se registró la misma cantidad de mediciones (58), sino que, asimismo, se removió la misma cantidad de datos atípicos (16). Lo que permitiría deducir que:

- Hay una fuerte distribución y correlación de los parámetros entre sí, por lo que cualquier afectación en uno u otro sugiere un cambio directo en los demás.
- Las técnicas de recolección y manipulación de muestras podrían haber tenido una falla de medición para todos los parámetros evaluados en un periodo de tiempo determinado.
- Los parámetros están afectados por aspectos geográficos o temporales.

TABLA 8.1: Datos atípicos eliminados por parámetros.

Parámetro	Tamaño original	Sin atípicos	Datos eliminados	Porcentaje
Arsénico	58	42	16	27,6 %
Sulfato	58	42	16	27,6 %
Clorofila-a	156	104	52	33,3 %
Cloruro	58	42	16	27,6 %
DO	355	270	85	23,9 %
Dureza	56	41	15	26,8 %
Fósforo	156	110	46	29,5 %

Parámetro	Tamaño original	Sin atípicos	Datos eliminados	Porcentaje
Nitrógeno	72	51	21	29,17 %
pH	414	376	38	9,18 %
Potasio	58	42	16	27,6 %
SDD	393	327	66	16,79 %
Sodio	58	42	16	27,6 %
Temperatura	415	377	38	9,16 %
TSS	78	53	25	32, 05 %
Turbidez	78	49	29	37,18 %

Fuente: Elaboración propia.

8.1.2. Selección de Variables Significativas.

Teniendo los marcos de datos sin valores atípicos, se procedió a seleccionar aquellas variables predictivas que mejor explicaban la variabilidad de las variables respuesta, es decir, de las observaciones de los parámetros. En esta parte del proceso se efectuó una metodología similar a la de *forward step-wise*, la cual consiste en la adición de variables al modelo, eligiendo aquella con el coeficiente de determinación múltiple (R^2) más alto (Bendel and Afifi, 1977). Dado que, hasta este punto, no solo se tenían las 12 bandas, sino que también se contaba con los diferentes índices normalizados junto con los coeficientes resultantes de la combinación de cada banda entre sí, no fue posible hacer una regresión con todas las variables porque esto supondría la obtención de problemas de multicolinealidad y sobreajuste, en la medida en que algunos parámetros iban a tener más variables predictivas que observaciones (Shrestha, 2020).

En consecuencia, se crearon modelos con una, dos y tres variables, se ajustaron las regresiones y se extrajo el R^2. Luego, se ordenaron los modelos resultantes de forma descendente y se eligieron los cinco (5) mejores. El procesamiento anteriormente descrito se implementó para cada uno de los parámetros de forma independiente. Asimismo, se aplicaron distintas transformaciones a la variable respuesta en búsqueda de un mejor ajuste.

A los modelos seleccionados se les aplicó un análisis de inflación de varianza (VIF), el cual es una técnica de detección de multicolinealidad basada en modelos de regresión lineal realizados sobre las variables independientes de la siguiente manera:

- Se toma una de las variables explicativas como variable dependiente.
- Se hace un modelo de regresión con las variables predictoras restantes.
- Se calcula el coeficiente de determinación R^2.
- Se calcula el VIF

$$VIF_i = \frac{1}{1 - R_i^2} \tag{8.6}$$

Donde VIF_i es el factor de inflación de varianza y R_i^2 es el coeficiente de determinación para la variable i (Thompson et al., 2017).

El índice VIF calculado se puede clasificar de la siguiente manera:

- VIF cercano a 1: multicolinealidad baja
- VIF menor o cercano a 5: multicolinealidad media
- VIF cercano o mayor a 10: multicolinealidad alta

Así entonces, con base en la ecuación (8.6), se eligió el modelo (dentro de los 5 preseleccionados) con el menor grado de multicolinealidad en sus variables. Durante este proceso, se pudo confirmar que la combinación de bandas unitarias entre sí genera una multicolinealidad muy alta, con índices VIF incluso mayores que 400, por tal motivo, se proyectó el ajuste de modelos con una combinación de índices normalizados y cocientes en lugar de bandas unitarias sin procesar.

8.1.3. Ajuste de los Modelos Preliminares.

Los modelos finales seleccionados se ajustaron utilizando la técnica de mínimos cuadrados ordinarios (MCO). Hasta este punto, ningún modelo se ha descartado por falta de ajuste; no obstante, los coeficientes de determinación obtenidos no permitían concluir que había un ajuste aceptable para ser considerado dentro del análisis. En ese orden de ideas, se eligieron tres (3) estadísticas como criterios de filtrado:

- Prueba F: la cual valida si el modelo se ajusta a los datos.

$$\begin{aligned} H_o &: \beta_1, \beta_2, ..., \beta_i = 0 \\ H_a &: \beta_1, \beta_2, ..., \beta_i \neq 0 \end{aligned} \tag{8.7}$$

En donde H_0 es la hipótesis nula, H_a es la hipótesis alterna y $\beta_1, \beta_2, ..., \beta_i$ son los coeficientes de regresión obtenidos en el ajuste. Si el p-valor es menor que el nivel de significancia α, se rechaza H_a y se concluye que los coeficientes son estadísticamente significativos.

- Prueba t: la cual valida la significancia estadística de cada coeficiente.

$$\begin{aligned} H_o &: \beta_i = 0 \\ H_a &: \beta_i \neq 0 \end{aligned} \tag{8.8}$$

En donde H_0 es la hipótesis nula, H_a es la hipótesis alternativa y β_i son los coeficientes de regresión obtenidos en el ajuste. Si el p-valor es menor que el nivel de significancia α, se rechaza H_a y se concluye que el coeficiente evaluado es estadísticamente significativo.

- Coeficiente de determinación múltiple R^2 ajustado: el cual toma en cuenta la inclusión de nuevas variables y su aporte a la explicación de la variabilidad de la variable respuesta.

8.1.4. Análisis de Observaciones Influyentes.

Debido a que el problema de falta de ajuste seguía presente en una cantidad considerable de modelos, se realizó un análisis de observaciones influyentes en el cálculo del coeficiente de determinación ajustado. La ecuación (8.9) ilustra las variables que intervienen en el cálculo. Nótese que la suma de cuadrados de los residuos tiene una relación inversa con el coeficiente; es decir, cuanto más grande es la suma, menor es el ajuste; por lo tanto, si una o más observaciones están muy distantes del valor

estimado, el ajuste disminuye sustancialmente. En ese orden de ideas, se obtuvieron los residuos de los modelos y se ordenaron de manera descendente; luego, se eliminaron las observaciones con los residuos más grandes. La implementación de esta metodología resultó en una mejora fundamental en el ajuste para la mayoría de los parámetros.

$$R^2_{ad} = 1 - \frac{(1 - \frac{SCE}{SCT})(N-1)}{N-p-1} \tag{8.9}$$

Donde R^2_{ad} es el coeficiente de determinación ajustado, SCE es la suma de cuadrados de los residuos, SCT es la suma total de cuadrados, N es el tamaño de la muestra y p es el número de variables independientes (Martínez, 2005).

La tabla (8.2) ilustra tanto los datos atípicos como los influyentes que fueron eliminados de las bases de datos consolidadas para todos los parámetros. El oxígeno disuelto (DO), por ejemplo, presentó un buen ajuste solamente hasta que se removió aproximadamente el 75 % de los datos, lo que sugiere que los datos recolectados por las estaciones en la zona de estudio no pueden ser explicados mediante la utilización de modelos de regresión lineal. Un fenómeno similar sucedió con el fósforo, cuyo comportamiento lineal no pudo ser explicado sino hasta la remoción del 71 % de las observaciones medidas. El bajo ajuste de estas variables no significa que no puedan ser modeladas; sino que, por el contrario, el enfoque de un modelamiento lineal podría no ser el adecuado según las características propias del parámetro en la zona de estudio, abriendo la posibilidad de explorar otras técnicas no lineales más avanzadas.

En términos generales, se podría deducir que el modelado mediante la utilización de regresiones lineales es bastante restrictivo en la medida en que es necesaria la eliminación de una cantidad representativa de datos que impacta directamente en la significancia y el ajuste general de los modelos; aún así, tras la implementación de las técnicas de limpieza y mejoramiento de ajuste, se logró obtener modelos más robustos y consistentes, capaces de ofrecer mayor confiabilidad en las estimaciones.

TABLA 8.2: Datos atípicos e influyentes eliminados en cada parámetros.

Parámetro	**Sin atípicos**	**Sin influyentes**	**Datos eliminados**	**Porcentaje total**
Arsénico	42	35	23	39,66 %
Sulfato	42	35	23	39,66 %
Clorofila-a	104	67	89	57,05 %
Cloruro	42	35	23	39,66 %
DO	270	95	260	73,24 %
Dureza	41	30	26	46,47 %
Fósforo	110	45	111	71,15 %
Nitrógeno	51	41	31	43,05 %
pH	376	241	173	41,79 %
Potasio	42	32	26	44,83 %
SDD	327	222	171	43,51 %
Sodio	42	33	25	43,10 %
Temperatura	377	207	258	62,17 %
TSS	53	39	39	50,00 %
Turbidez	49	38	40	51,28 %

Fuente: Elaboración propia.

8.2. Ajuste de los Modelos Finales.

Una vez que se obtuvieron los coeficientes de determinación ajustados aceptables, se procedió a validar el resto de los supuestos:

8.2.1. Arsénico.

El resumen de los resultados de la regresión lineal sobre los datos del arsénico se presenta en la tabla (8.3). Este parámetro se modeló con un total de 35 datos, arrojando un R^2 de 0,849 y un R^2 ajustado de 0,834. En el proceso, se probó una combinación de transformaciones, resultando en la cúbica como la que mejor ajuste presentó. La probabilidad del estadístico F fue muy cercana a 0, por lo que se puede concluir que el modelo se ajusta a los datos. La probabilidad de la prueba t para los parámetros está muy cercana a 0; por consiguiente, se concluye que los coeficientes son altamente significativos; no obstante, el valor de la prueba de *Durbin-Watson* (0,729) alude a una autocorrelación positiva; es decir, si un error es grande en una observación, el siguiente tiende a ser igualmente grande. Este tipo de problemas sucede cuando los datos objeto de análisis tienen componentes temporales o geográficos, estando ambos aspectos presentes en la información obtenida.

TABLA 8.3: Resumen del modelo de regresión lineal por MCO para el arsénico.

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.849
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	57.93
Date:	Thu, 13 Nov 2025	Prob (F-statistic):	8.22e-13
No. Observations:	35	AIC:	-83.72
Df Residuals:	31	BIC:	-77.50

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3039	0.014	22.232	0.000	0.276	0.332
ND_b8a_b9	0.0066	0.001	8.576	0.000	0.005	0.008
ND_b1_b5	0.3341	0.062	5.346	0.000	0.207	0.462
ND_b9_b12	-0.0150	0.002	-7.331	0.000	-0.019	-0.011

Omnibus:	0.608	Durbin-Watson:	0.729
Prob(Omnibus):	0.738	Jarque-Bera (JB):	0.659
Skew:	0.030	Prob(JB):	0.719
Kurtosis:	2.330	Cond. No.	91.5

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

La regresión con MCO, en consecuencia, no fue la opción adecuada para modelar el parámetro del arsénico. Por lo tanto, se ajustó un modelo mediante mínimos cuadrados generalizados con autocorrelación o GLSAR (*Generalized Least Squares Autoregressive*, por sus siglas en inglés). Este tipo de regresión no solo corrige problemas de autocorrelación, sino que también puede ser útil para resolver la heterocedasticidad. La tabla (8.4) muestra que, si bien el ajuste general del modelo disminuyó (R^2

ajustado de 0,747), tanto la significancia general del modelo como la de los estimadores se mantuvieron. Finalmente, el valor de la prueba de *Durbin-Watson* está dentro del rango permitido para la validación del supuesto en cuestión.

TABLA 8.4: Resumen del modelo de regresión lineal por GLSAR para el arsénico.

```
==============================================================================
                         GLSAR Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.770
Model:                          GLSAR   Adj. R-squared:                  0.747
Method:               Least Squares   F-statistic:                     168.1
Date:              Thu, 13 Nov 2025   Prob (F-statistic):           7.61e-19
No. Observations:                  34   AIC:                            -99.97
Df Residuals:                      30   BIC:                            -93.86
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3069      0.033      9.231      0.000       0.239       0.375
ND_b8a_b9      0.0062      0.000     20.777      0.000       0.006       0.007
ND_b1_b5       0.2488      0.060      4.179      0.000       0.127       0.370
ND_b9_b12     -0.0135      0.001     -9.831      0.000      -0.016      -0.011
==============================================================================
Omnibus:                        5.490   Durbin-Watson:                   1.634
Prob(Omnibus):                  0.064   Jarque-Bera (JB):                6.445
Skew:                           0.111   Prob(JB):                       0.0399
Kurtosis:                       5.121   Cond. No.                         79.5
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Con respecto a los supuestos de normalidad y homocedasticidad. Se ejecutaron las pruebas de *Shapiro-Wilk*, *Kolmogorov-Smirnov* y *Jarque-Bera*, las cuales arrojaron p-valores de 0,0806, 0,489 y 0,4249, respectivamente, por lo que no se rechaza la normalidad en los residuos. Debido a que la prueba de heterocedasticidad de *Breusch-Pagan* arrojó un p-valor de 0,091, se ajustó el modelo con la estructura de covarianza 'HAC' (*Heteroskedasticity and Autocorrelation Consistent*, por sus siglas en inglés), la cual robustece los errores estándar en muestras con problemas de heterocedasticidad y autocorrelación (Andrews, 1991). Por último, para la validación final del modelo, se utilizó el método *leave-one-out*, el cual realiza regresiones con los datos finales dejando uno fuera, repitiendo el proceso hasta descartar la última observación, computando el error medio absoluto y el error medio cuadrático, además del coeficiente de determinación.

La validación cruzada para el arsénico mostró que tanto el error medio absoluto (MAE) como el error medio cuadrático (RMSE) son consistentes y bajos, con valores de 0,0605 y 0,0716, respectivamente, por lo que las estimaciones con la estructura HAC son estables y pueden considerarse predicciones confiables. En cuanto al coeficiente de determinación, el resultado final indica que la variabilidad es explicada en más de un 80 %, un valor sólido que indica confiabilidad y robustez en el ajuste. En conclusión, el modelo GLSAR para el parámetro del arsénico cumplió con los supuestos clásicos de la regresión lineal y la validación cruzada ofreció confiabilidad al momento de predecir.

8.2.2. Sulfato.

El resumen del modelo para el sulfato con MCO se presenta en la tabla (8.5). Como era de esperarse, se encontraron resultados similares a los del caso de arsénico, con un buen R^2 ajustado de 0,87 y una probabilidad del estadístico tanto de la prueba F como de las pruebas t muy cercana a 0. No obstante, el supuesto de la autocorrelación no se cumplió, debido a que el valor de la prueba de *Durbin-Watson* está muy lejano de 2; por lo tanto, se pudo concluir que los residuos tienen correlación positiva. La variable respuesta para este parámetro también fue objeto de transformación para la obtención de un mejor ajuste y el cumplimiento del supuesto de no heterocedasticidad. La transformación en mención fue la logarítmica.

TABLA 8.5: Resumen del modelo de regresión lineal por MCO para el sulfato.

```
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.878
Model:                            OLS   Adj. R-squared:                  0.870
Method:                 Least Squares   F-statistic:                     115.2
Date:                Sun, 16 Nov 2025   Prob (F-statistic):           2.41e-15
No. Observations:                  35   AIC:                            -32.67
Df Residuals:                      32   BIC:                            -28.01
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.0245      0.082     37.033      0.000       2.858       3.191
ND_b5_b11      3.1233      0.207     15.091      0.000       2.702       3.545
ND_b8_b11     -2.5784      0.200    -12.883      0.000      -2.986      -2.171
==============================================================================
Omnibus:                        0.084   Durbin-Watson:                   1.186
Prob(Omnibus):                  0.959   Jarque-Bera (JB):                0.088
Skew:                           0.078   Prob(JB):                        0.957
Kurtosis:                       2.811   Cond. No.                         14.2
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

La tabla (8.6) muestra el resumen del ajuste con GLSAR. Si bien el ajuste general del modelo disminuyó un poco, el hecho de que la variabilidad del sulfato se explique en un 75 % no determina una caída sustancial y se considera aceptable. Los otros parámetros que no se vieron afectados fueron el estadístico de la prueba F y los de las pruebas t; por lo tanto, se cumplió la significancia estadística tanto del modelo como de los estimadores, incluso para un nivel del 1 %. El valor de la prueba de *Durbin-Watson* de 1,707 indicó que el problema de la autocorrelación fue corregido satisfactoriamente.

Las pruebas de normalidad y homocedasticidad fueron satisfactorias al arrojar p-valores de 0,1085, 0,739 y 0,8247 para las pruebas de *Shapiro-Wilk*, *Kolmogorov-Smirnov* y *Jarque-Bera*, respectivamente. Si bien la prueba de *Goldfeld-Quandt* arrojó un p-valor de 0,0389 para la evaluación de la heterocedasticidad, las pruebas de *Breusch-Pagan* y *White* produjeron p-valores de 0,4318 y 0,4335, respectivamente. Por lo tanto, no se halló evidencia para rechazar la homocedasticidad en las varianzas de los residuos.

TABLA 8.6: Resumen del modelo de regresión lineal por GLSAR para el sulfato.

```
==============================================================================
                         GLSAR Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.768
Model:                          GLSAR   Adj. R-squared:                  0.753
Method:                 Least Squares   F-statistic:                     21.06
Date:                Sun, 16 Nov 2025   Prob (F-statistic):           1.67e-06
No. Observations:                  34   AIC:                            -37.84
Df Residuals:                      31   BIC:                            -33.26
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.0966      0.113     27.300      0.000       2.865       3.328
ND_b5_b11      2.8283      0.437      6.476      0.000       1.938       3.719
ND_b8_b11     -2.3679      0.403     -5.876      0.000      -3.190      -1.546
==============================================================================
Omnibus:                        8.653   Durbin-Watson:                   1.707
Prob(Omnibus):                  0.013   Jarque-Bera (JB):               10.831
Skew:                           0.572   Prob(JB):                      0.00445
Kurtosis:                       5.518   Cond. No.                         10.9
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Debido a la cantidad de datos disponibles, se realizó la validación cruzada del tipo *leave-one-out*. Al igual que con el arsénico, tanto el error medio absoluto (0,1216) como el error medio cuadrático (0,1537) fueron bajos. El coeficiente de determinación de la validación fue de 0,8514, lo que indica que el modelo tiene un buen poder de predicción y generalización ante ligeras modificaciones en la base de datos con la que se ajustó inicialmente. En conclusión, los supuestos clásicos de la regresión lineal se cumplieron satisfactoriamente y la validación cruzada ofrece confiabilidad y robustez en el ajuste final.

8.2.3. Clorofila-a.

El ajuste inicial del modelo por mínimos cuadrados ordinarios, que se presenta en la tabla (8.7), proporcionó un gran resultado con un 87 % de la variabilidad explicada por las variables independientes; sin embargo, es pertinente hacer la aclaración de que este parámetro en específico fue uno de los más complejos de modelar, debido a que los valores atípicos e influyentes afectaron directamente el ajuste. Esto ocasionó que una gran cantidad de datos fuera removida para permitir que el modelo se ajustara a los datos. La significancia general del modelo y de los estimadores cumple con el nivel de significancia de hasta el 1 %. No obstante, los residuos de este parámetro también están fuertemente autocorrelacionados; por lo tanto, se aplicó nuevamente el modelo GLSAR para corregir dicho problema. Asimismo, la transformación logarítmica se usó en este parámetro, lo cual permitió resolver problemas de heterocedasticidad y falta de ajuste.

TABLA 8.7: Resumen del modelo de regresión lineal por MCO para la clorofila.

OLS Regression Results			
Dep. Variable:	Y	R-squared:	0.884
Model:	OLS	Adj. R-squared:	0.881
Method:	Least Squares	F-statistic:	244.4
Date:	Mon, 08 Dec 2025	Prob (F-statistic):	1.09e-30
No. Observations:	67	AIC:	-4.941
Df Residuals:	64	BIC:	1.673

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9688	0.035	85.147	0.000	2.899	3.038
ND_b2_b7	-2.7978	0.195	-14.370	0.000	-3.187	-2.409
ND_b3_b8	3.2087	0.154	20.858	0.000	2.901	3.516

Omnibus:	47.066	Durbin-Watson:	1.056
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5.670
Skew:	-0.114	Prob(JB):	0.0587
Kurtosis:	1.593	Cond. No.	8.98

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Hasta este punto, y con los resultados de la tabla (8.8), se puede concluir que la regresión por mínimos cuadrados generalizados castiga notablemente el ajuste, haciendo que el coeficiente de determinación ajustado disminuya. Para el caso de la clorofila-a, el R^2 ajustado pasó de 87 % a 75,3 %, prácticamente una disminución de 12 puntos. Si bien es una caída sustancial en la capacidad predictiva del modelo, el valor del estadístico de *Durbin-Watson* de 2,202 garantizó que la autocorrelación en los residuos ya no estuviera presente, brindando confiabilidad en los resultados. Adicionalmente, los estimadores y el modelo general cumplieron con el nivel de significancia.

La normalidad en los residuos se validó con la prueba de *Jarque-Bera*, la cual arrojó un p-valor de 0,0626. Si bien se cumplió el supuesto, al estar dicho resultado tan cerca del nivel de significancia, fue necesario confirmarlo con la prueba de *Kolmogorov-Smirnov*, cuyo p-valor fue de 0,127. Por lo tanto,

no hubo evidencia para rechazar la normalidad en los residuos. De igual forma, la probabilidad de la prueba de *Goldfeld-Quandt* de 0,2979 indicó que tampoco hay evidencia para rechazar la homocedasticidad.

TABLA 8.8: Resumen del modelo de regresión lineal por GLSAR para la clorofila.

GLSAR Regression Results			
Dep. Variable:	Y	R-squared:	0.753
Model:	GLSAR	Adj. R-squared:	0.745
Method:	Least Squares	F-statistic:	146.3
Date:	Mon, 08 Dec 2025	Prob (F-statistic):	2.12e-24
No. Observations:	66	AIC:	-22.42
Df Residuals:	63	BIC:	-15.85

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9552	0.041	72.303	0.000	2.874	3.037
ND_b2_b7	-2.6177	0.294	-8.918	0.000	-3.204	-2.031
ND_b3_b8	2.9818	0.246	12.121	0.000	2.490	3.473

Omnibus:	9.099	Durbin-Watson:	2.202
Prob(Omnibus):	0.011	Jarque-Bera (JB):	3.872
Skew:	-0.322	Prob(JB):	0.144
Kurtosis:	2.003	Cond. No.	7.09

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Finalmente, los resultados de la validación cruzada indicaron que el modelo, además de cumplir con los supuestos clásicos, también presentó un buen poder predictivo y ofreció confiabilidad y robustez en las estimaciones, al obtenerse un error medio absoluto de 0,2144 y un error medio cuadrático de 0,237, a pesar de haber tenido una reducción considerable en el coeficiente de determinación ajustado. Como ya se mencionó anteriormente, el hecho de que el MAE y el RMSE fueran similares demuestra la estabilidad del modelo ante cambios en la muestra de partida.

8.2.4. Cloruro.

Siguiendo la línea de los parámetros relacionados con minerales, el cloruro presentó un ajuste de 0,861, dado por el coeficiente de determinación ajustado; aun así, el problema de la autocorrelación estuvo presente en este modelo también. Teniendo esto en cuenta, es importante mencionar que el modelo fue estadísticamente significativo, respaldado por la probabilidad de la prueba F, al igual que cada uno de los estimadores, cuyas probabilidades de las pruebas t estuvieron muy cercanas a 0. La transformación logarítmica también se utilizó en esta variable, permitiendo elevar el R^2 ajustado en comparación con el calculado con la variable sin transformar. Los resultados completos de la regresión se presentan en la tabla (8.9).

TABLA 8.9: Resumen del modelo de regresión lineal por MCO para el cloruro.

OLS Regression Results			
Dep. Variable:	Y	R-squared:	0.869
Model:	OLS	Adj. R-squared:	0.861
Method:	Least Squares	F-statistic:	106.2
Date:	Thu, 13 Nov 2025	Prob (F-statistic):	7.50e-15
No. Observations:	35	AIC:	-45.80
Df Residuals:	32	BIC:	-41.13

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.5317	0.064	70.992	0.000	4.402	4.662
ND_b5_b11	2.5230	0.173	14.567	0.000	2.170	2.876
ND_b8_b11	-1.9542	0.163	-11.987	0.000	-2.286	-1.622

Omnibus:	4.004	Durbin-Watson:	1.266
Prob(Omnibus):	0.135	Jarque-Bera (JB):	2.002
Skew:	-0.289	Prob(JB):	0.368
Kurtosis:	1.981	Cond. No.	14.0

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Los resultados de la regresión por mínimos cuadrados generalizados, que se muestran en la tabla (8.10), resolvieron el problema de la autocorrelación para el cloruro de manera satisfactoria, sin afectar ostensiblemente la significancia estadística, tanto del modelo como de los estimadores. Una vez más, se repitió el fenómeno de la reducción en el ajuste; aún así, que las variables predictoras expliquen un 74,7 % de la variabilidad del cloruro sigue considerándose aceptable para un modelo con solamente 34 observaciones. Adicionalmente, los errores estándar se calcularon con la estructura HAC para asegurar la integridad de los intervalos de confianza.

Para la validación de la normalidad en los residuos, se utilizó la prueba de *Shapiro-Wilk*, la cual arrojó un p-valor de 0,0883. Debido a que la probabilidad mencionada anteriormente estuvo muy cerca del nivel de significancia, se adicionaron las pruebas de *Kolmogorov-Smirnov* y *Jarque-Bera*, cuyas probabilidades fueron de 0,5884 y 0,5261, respectivamente; por lo tanto, se considera que los residuos tienen una distribución normal. En cuanto al supuesto de no heterocedasticidad, tanto la prueba

TABLA 8.10: Resumen del modelo de regresión lineal por GLSAR para el cloruro.

```
==============================================================================
                         GLSAR Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.762
Model:                          GLSAR   Adj. R-squared:                  0.747
Method:                 Least Squares   F-statistic:                     22.97
Date:                Thu, 13 Nov 2025   Prob (F-statistic):           7.61e-07
No. Observations:                  34   AIC:                            -49.62
Df Residuals:                      31   BIC:                            -45.04
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.5992      0.110     41.975      0.000       4.376       4.823
ND_b5_b11      2.2867      0.337      6.777      0.000       1.599       2.975
ND_b8_b11     -1.8139      0.299     -6.066      0.000      -2.424      -1.204
==============================================================================
Omnibus:                        1.914   Durbin-Watson:                   1.658
Prob(Omnibus):                  0.384   Jarque-Bera (JB):                0.855
Skew:                          -0.198   Prob(JB):                        0.652
Kurtosis:                       3.669   Cond. No.                         11.2
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

de *White* como la de *Breusch-Pagan* produjeron p-valores notablemente más grandes que el nivel de significancia, por lo que se valida la varianza constante a lo largo de la muestra.

Resultados similares se obtuvieron para la validación cruzada de este modelo, logrando un error medio absoluto de 0,1057 y un error medio cuadrático de 0,1275. Esto indicó que el modelo tuvo una buena capacidad de predicción al tener errores pequeños. Adicionalmente, el coeficiente R^2 demostró, una vez más, la correcta generalización de la regresión.

8.2.5. Oxígeno disuelto.

La regresión por MCO se muestra en la tabla (8.11). Analizando el coeficiente R^2, este parámetro fue el que mejor ajuste presentó al explicar más del 95 % de la variabilidad de las observaciones; tanto los estimadores como el modelo son estadísticamente significativos, respaldados por las bajas probabilidades obtenidas de sus correspondientes pruebas. No obstante, el valor del estadístico de *Durbin-Watson* estuvo notablemente alejado de 2. A priori, el oxígeno disuelto pareciera ser un parámetro con problemas sencillos de resolver; no obstante, para obtener un ajuste decente fue necesario eliminar cerca del 75 % de los datos, además de la transformación de la variable dependiente con el logaritmo natural. Lo que esto conlleva es que la robustez del modelo se vea directamente comprometida y que la capacidad predictiva del modelo disminuya al tener que forzarlo a una linealidad inexistente.

TABLA 8.11: Resumen del modelo de regresión lineal por MCO para el oxígeno disuelto.

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.954
Model:	OLS	Adj. R-squared:	0.953
Method:	Least Squares	F-statistic:	964.7
Date:	Mon, 08 Dec 2025	Prob (F-statistic):	1.88e-62
No. Observations:	95	AIC:	-491.9
Df Residuals:	92	BIC:	-484.2

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.2592	0.006	387.859	0.000	2.248	2.271
ND_b3_b9	-0.0533	0.002	-29.706	0.000	-0.057	-0.050
b7	-15.8837	0.376	-42.280	0.000	-16.630	-15.138

Omnibus:	18.093	Durbin-Watson:	0.370
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5.031
Skew:	-0.178	Prob(JB):	0.0808
Kurtosis:	1.930	Cond. No.	332.

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

A pesar de lo mencionado anteriormente, se ajustó el modelo con GSLAR, como se evidencia en la tabla (8.11). El ajuste general disminuyó más de 20 puntos, obteniéndose un R^2 ajustado de 0,729 que, si bien no fue despreciable, sí evidencia la fragilidad de los datos ante cambios en las condiciones de la regresión. La autocorrelación fue resuelta dado un valor de 1,749 para *Durbin-Watson*; sin embargo, el condicionamiento de la eliminación masiva de datos fue una restricción que generó una alerta bastante relevante al momento de generar estimaciones con el modelo.

Las pruebas de normalidad, por otra parte, fallaron categóricamente; tanto *Shapiro-Wilk* como *Jarque-Bera* y *Kolmogorov-Smirnov* generaron p-valores por debajo del nivel de significancia del 5 %. Si bien el modelo arrojó resultados satisfactorios para las pruebas de homocedasticidad en las varianzas de los residuos de *Breusch-Pagan* y *Goldfeld-Quandt*, también falló en la prueba de *White*, conocida por ser una de las más robustas. En resumen, el modelo no cumplió con dos de los supuestos clásicos

TABLA 8.12: Resumen del modelo de regresión lineal por GLSAR para el oxígeno disuelto.

```
==============================================================================
                         GLSAR Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.735
Model:                          GLSAR   Adj. R-squared:                  0.729
Method:                 Least Squares   F-statistic:                     69.60
Date:                Mon, 08 Dec 2025   Prob (F-statistic):           4.58e-19
No. Observations:                  94   AIC:                            -659.7
Df Residuals:                      91   BIC:                            -652.0
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1894      0.020    108.084      0.000       2.149       2.230
ND_b3_b9      -0.0330      0.003    -11.665      0.000      -0.039      -0.027
b7            -9.4752      0.873    -10.848      0.000     -11.210      -7.740
==============================================================================
Omnibus:                       51.542   Durbin-Watson:                   1.749
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              190.315
Skew:                           1.821   Prob(JB):                     4.72e-42
Kurtosis:                       8.944   Cond. No.                         802.
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

de la regresión lineal; además, para lograr un ajuste decente por GSLAR, fue necesaria la eliminación de casi toda la muestra. Por tal motivo, este parámetro se descartó y no se ejecutó ningún tipo de validación cruzada.

8.2.6. Dureza.

La dureza se modeló inicialmente con un total de 30 datos y dos variables independientes, obteniéndose un coeficiente de determinación ajustado de 0,858. El estadístico F para el modelo presentó un valor alto, por lo que se concluye que el modelo se ajusta a los datos y que, además, los estimadores son significativos al 1 %. Como ha sido la tendencia, el problema de la autocorrelación positiva volvió a estar presente en este parámetro, siguiendo el comportamiento de sus similares relacionados con minerales. Adicionalmente, las observaciones de la dureza arrojaron resultados preliminares satisfactorios en cuanto a homocedasticidad y ajuste; por lo tanto, no fue requerida ninguna transformación. El resumen completo de los resultados se presenta en la tabla (8.13).

TABLA 8.13: Resumen del modelo de regresión lineal por MCO para la dureza.

```
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.867
Model:                            OLS   Adj. R-squared:                  0.858
Method:                 Least Squares   F-statistic:                     88.35
Date:                Mon, 08 Dec 2025   Prob (F-statistic):           1.42e-12
No. Observations:                  30   AIC:                              242.8
Df Residuals:                      27   BIC:                              247.0
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         90.2658      7.493     12.047      0.000      74.892     105.640
ND_b5_b11    252.7933     19.075     13.253      0.000     213.655     291.931
ND_b8_b11   -203.2807     18.314    -11.100      0.000    -240.857    -165.704
==============================================================================
Omnibus:                        5.818   Durbin-Watson:                   1.028
Prob(Omnibus):                  0.055   Jarque-Bera (JB):                2.111
Skew:                           0.247   Prob(JB):                        0.348
Kurtosis:                       1.798   Cond. No.                         13.3
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

El modelo GLSAR sobre la muestra de dureza se presenta en la tabla (8.14), el cual indicó que el R^2 ajustado disminuyó; no obstante, logra explicar la variabilidad de la variable independiente en un 73 %. La autocorrelación fue corregida apropiadamente, soportada por un valor de *Durbin-Watson* de 1,896. Las pruebas de normalidad utilizadas fueron las de *Shapiro-Wilk*, *Kolmogorov-Smirnov* y *Jarque-Bera*, las cuales generaron probabilidades bastante altas; en consecuencia, hay evidencia sólida para concluir que los residuos siguen una distribución normal. La homocedasticidad en la varianza de los residuos también fue objeto de no rechazo al pasar las pruebas de *Breusch-Pagan* y *Goldfeld-Quandt* con probabilidades de 0,4383 y 0,9145, respectivamente.

Finalmente, con el cumplimiento de todos los supuestos clásicos de la regresión lineal, se procedió a ejecutar una rutina de validación cruzada tipo *leave-one-out*, la cual produjo un error medio absoluto de 0,0926, un error medio cuadrático de 0,1151 y un R^2 ajustado de 0,8108. En este orden de ideas,

se pudo evidenciar una consistencia y robustez en el modelo, permitiendo generar confiabilidad y estabilidad en las estimaciones.

TABLA 8.14: Resumen del modelo de regresión lineal por GLSAR para la dureza.

```
==============================================================================
                         GLSAR Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.750
Model:                           GLSAR   Adj. R-squared:                  0.731
Method:                 Least Squares   F-statistic:                     33.07
Date:                Mon, 08 Dec 2025   Prob (F-statistic):           7.19e-08
No. Observations:                  29   AIC:                             223.8
Df Residuals:                      26   BIC:                             227.9
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         97.7495      8.644     11.309      0.000      79.982     115.517
ND_b5_b11    217.5331     26.847      8.103      0.000     162.347     272.719
ND_b8_b11   -178.3453     24.665     -7.231      0.000    -229.044    -127.646
==============================================================================
Omnibus:                        1.734   Durbin-Watson:                   1.896
Prob(Omnibus):                  0.420   Jarque-Bera (JB):                1.442
Skew:                           0.526   Prob(JB):                        0.486
Kurtosis:                       2.704   Cond. No.                         9.92
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

8.2.7. Fósforo.

El fósforo, al igual que en el caso del oxígeno disuelto, presentó varias irregularidades al momento de ajustar el modelo por mínimos cuadrados ordinarios. En primer lugar, si bien arrojó un coeficiente de correlación ajustado más que decente de 0,851, para obtener este resultado fue necesario remover más del 70 % de la muestra, ya que tanto los valores atípicos como los influyentes representaron un impacto importante.

En otras palabras, la muestra inicial contenía un total de 156 datos para el periodo en cuestión; se removieron un total de 46 datos atípicos y, posteriormente, se eliminaron 111 datos adicionales del análisis con el fin de mejorar el ajuste mediante la evaluación de residuos influyentes. Finalmente, el modelo se ajustó con un total de 45 datos restantes y dos variables independientes. Pese a esto, el valor de la probabilidad del estadístico F para la significancia del modelo sugirió que el modelo se ajusta a los datos; adicionalmente, la significancia estadística de los estimadores también se validó con los p-valores de las pruebas t correspondientes. El resumen de la regresión por MCO se presenta en la tabla (8.15).

TABLA 8.15: Resumen del modelo de regresión lineal por MCO para el fósforo.

```
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.858
Model:                            OLS   Adj. R-squared:                  0.851
Method:                 Least Squares   F-statistic:                     126.9
Date:                Mon, 08 Dec 2025   Prob (F-statistic):           1.58e-18
No. Observations:                  45   AIC:                            -251.9
Df Residuals:                      42   BIC:                            -246.5
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0825      0.002     34.145      0.000       0.078       0.087
ND_b4_b8      -0.1291      0.009    -13.621      0.000      -0.148      -0.110
ND_b9_b11      0.0022      0.000      6.877      0.000       0.002       0.003
==============================================================================
Omnibus:                       10.074   Durbin-Watson:                   0.908
Prob(Omnibus):                  0.006   Jarque-Bera (JB):                3.100
Skew:                          -0.248   Prob(JB):                        0.212
Kurtosis:                       1.814   Cond. No.                         30.7
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

El problema de enfoque relacionado con el ajuste lineal del parámetro del fósforo se corroboró con el modelo por mínimos cuadrados generalizados, técnica que fue de gran utilidad para resolver el problema de la autocorrelación en los parámetros mencionados hasta ahora, en la medida en que el ajuste se desplomó drásticamente al presentar un coeficiente de correlación ajustado de tan solo 0,179 tras haber sido de 0,851 por MCO. Paralelamente, el valor del estadístico de la prueba F disminuyó proporcionalmente, y los estimadores pasaron a ser no significativos a un nivel de confianza del 1 %. Los resultados completos se muestran en la tabla (8.16).

TABLA 8.16: Resumen del modelo de regresión lineal por GLSAR para el fósforo.

GLSAR Regression Results			
Dep. Variable:	Y	R-squared:	0.217
Model:	GLSAR	Adj. R-squared:	0.179
Method:	Least Squares	F-statistic:	5.695
Date:	Fri, 14 Nov 2025	Prob (F-statistic):	0.00657
No. Observations:	44	AIC:	-326.4
Df Residuals:	41	BIC:	-321.0

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1001	0.010	9.822	0.000	0.080	0.121
ND_b4_b8	-0.0278	0.008	-3.326	0.002	-0.045	-0.011
ND_b9_b11	0.0004	0.000	2.626	0.012	8.87e-05	0.001

Omnibus:	18.784	Durbin-Watson:	1.734
Prob(Omnibus):	0.000	Jarque-Bera (JB):	25.589
Skew:	1.366	Prob(JB):	2.78e-06
Kurtosis:	5.548	Cond. No.	93.6

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Si bien el problema de la autocorrelación se solucionó, el modelo presentó muchos más problemas que impidieron validar algún tipo de confiabilidad, robustez o integridad en las estimaciones. Por ejemplo, el supuesto de normalidad tampoco se cumplió, ya que la probabilidad de la prueba de *Shapiro-Wilk* fue de 0,0091. Por otra parte, no se halló evidencia para rechazar la homocedasticidad en las varianzas de la muestra de los residuos, por lo que el modelo no cumplió con dos de los cuatro supuestos clásicos de la regresión lineal. En resumen, por los motivos mencionados anteriormente, el fósforo no se tuvo en cuenta en el análisis posterior.

8.2.8. Nitrógeno.

El nitrógeno, desde el inicio, presentó resultados poco confiables. En primer lugar, el ajuste fue bastante pobre, con un coeficiente de determinación ajustado de 0,547 y un total de 41 observaciones. Aun así, las pruebas de significancia del modelo y de los estimadores fueron satisfactorias a un nivel de significancia del 1 %. Por otra parte, el valor de la prueba para la detección de autocorrelación en los residuos de *Durbin-Watson* estuvo considerablemente alejado de 2. Dado los resultados presentados en la tabla (8.17), se ejecutó el modelo por mínimos cuadrados generalizados en búsqueda de un mejor ajuste.

El modelo GLSAR, en este caso, no fue de gran ayuda. El ajuste fue pésimo al arrojar un R^2 de 0,031. Por primera vez, el modelo falló la prueba de significancia estadística de la prueba F, con una probabilidad de 0,558. De la misma manera, los estimadores dejaron de ser significativos con p-valores de la prueba t de 0,294 y 0,304 para el índice de la banda 5 con la banda 8a y el índice de la banda 8 con la banda 9, respectivamente. Además de lo mencionado anteriormente, se eliminaron datos cuyos residuales fueron mayores, como se ha venido haciendo con el análisis de las observaciones influyentes; a pesar de esto, el ajuste con MCO se estancó luego de la eliminación de más del 60 % de la muestra y, a su vez, el modelo GLSAR tampoco tendió a mejorar con esta técnica.

En síntesis, las observaciones recolectadas para el parámetro del nitrógeno demostraron que la regresión lineal no es un buen enfoque para su modelación; por lo tanto, este mineral también fue descartado del análisis al no cumplir con el ajuste mínimo para considerarse aceptable. Los supuestos clásicos de normalidad y homocedasticidad no se evaluaron por los motivos ya conocidos.

TABLA 8.17: Resumen del modelo de regresión lineal por MCO para el nitrógeno.

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.570
Model:	OLS	Adj. R-squared:	0.547
Method:	Least Squares	F-statistic:	25.15
Date:	Fri, 14 Nov 2025	Prob (F-statistic):	1.10e-07
No. Observations:	41	AIC:	-90.91
Df Residuals:	38	BIC:	-85.77

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4421	0.029	15.405	0.000	0.384	0.500
ND_b5_b8a	-0.6505	0.103	-6.326	0.000	-0.859	-0.442
ND_b8_b9	0.0163	0.002	6.660	0.000	0.011	0.021

Omnibus:	5.707	Durbin-Watson:	0.751
Prob(Omnibus):	0.058	Jarque-Bera (JB):	2.037
Skew:	-0.021	Prob(JB):	0.361
Kurtosis:	1.909	Cond. No.	59.5

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

8.2.9. Acidez (pH).

La acidez fue uno de los parámetros con la mayor cantidad de observaciones, con 414 muestras recolectadas durante el periodo efectivo del proyecto. Finalmente, luego de la fase de limpieza de datos atípicos e influyentes, se ajustó el modelo por MCO con 241 datos, es decir, con el 42 % de la muestra, aproximadamente. El ajuste inicial fue de 0,896 para el R^2 y de 0,895 para el R^2 ajustado. El hecho de que estos dos valores estuvieran tan próximos el uno al otro indica que las variables independientes seleccionadas inciden fuertemente en el ajuste del modelo. El valor de la prueba del estadístico F fue notablemente alto, confirmando que el modelo se ajusta a los datos. Asimismo, los valores de la prueba t para los estimadores arrojaron resultados confiables, por lo que las variables fueron estadísticamente significativas a un nivel de confianza del 1 %. No obstante, fue posible evidenciar problemas con relación a los supuestos de autocorrelación y normalidad al presentar un valor de 0,927 en la prueba de *Durbin-Watson* y una probabilidad de 0,0028 en la prueba de *Jarque-Bera*. Los resultados completos se muestran en la tabla (8.18).

TABLA 8.18: Resumen del modelo de regresión lineal por MCO para el pH.

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.896
Model:	OLS	Adj. R-squared:	0.895
Method:	Least Squares	F-statistic:	1025.
Date:	Fri, 14 Nov 2025	Prob (F-statistic):	1.09e-117
No. Observations:	241	AIC:	-254.2
Df Residuals:	238	BIC:	-243.7

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.0861	0.021	340.063	0.000	7.045	7.127
ND_b5_b8a	1.4270	0.054	26.301	0.000	1.320	1.534
b1	22.5928	0.692	32.646	0.000	21.229	23.956

Omnibus:	32.881	Durbin-Watson:	0.927
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11.725
Skew:	-0.283	Prob(JB):	0.00284
Kurtosis:	2.080	Cond. No.	78.3

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Se obtuvo una disminución en el ajuste general del modelo por mínimos cuadrados generalizados, como se muestra en la tabla (8.19); no obstante, que la variabilidad de la acidez sea explicada en un 74,6 % por las variables explicativas sigue considerándose aceptable. Si bien los valores de los estadísticos de la prueba F y de las pruebas t disminuyeron, no hay evidencia suficiente para determinar que el modelo no se ajusta a los datos y que los estimadores no son estadísticamente significativos. La autocorrelación se corrigió con el modelo GLSAR, con un valor de 2,262 para la prueba de *Durbin-Watson*.

El supuesto de normalidad fue rechazado, inicialmente, por la prueba de *Jarque-Bera*; sin embargo, se realizó una segunda con el método de *Kolmogorov-Smirnov*, arrojando un valor de 0,1077. Si bien es un valor bajo y teniendo en cuenta tanto el valor de la asimetría como el de la curtosis, se consideró que los residuos son aproximadamente normales. Para la homocedasticidad, se eligió la prueba de *Goldfeld-Quandt*, la cual produjo una probabilidad de 0,976; por tanto, se concluye que la varianza en los residuos es constante a lo largo de la muestra.

Por último, la validación cruzada arrojó un coeficiente de correlación de 0,8927, un error medio absoluto de 0,12 y un error medio cuadrático de 0,1432. Esto indicó que el modelo fue consistente a lo largo de la muestra, ofreciendo confiabilidad en las estimaciones. En conclusión, salvo la normalidad, todos los demás supuestos clásicos se cumplieron satisfactoriamente y el ajuste obtenido permitió considerar el modelo apto para predecir.

TABLA 8.19: Resumen del modelo de regresión lineal por GLSAR para el pH.

GLSAR Regression Results

Dep. Variable:	Y	R-squared:	0.749
Model:	GLSAR	Adj. R-squared:	0.746
Method:	Least Squares	F-statistic:	256.0
Date:	Fri, 14 Nov 2025	Prob (F-statistic):	6.13e-60
No. Observations:	240	AIC:	-334.0
Df Residuals:	237	BIC:	-323.6

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.0804	0.045	156.496	0.000	6.991	7.170
ND_b5_b8a	1.4851	0.102	14.543	0.000	1.284	1.686
b1	22.1631	1.452	15.268	0.000	19.303	25.023

Omnibus:	10.668	Durbin-Watson:	2.262
Prob(Omnibus):	0.005	Jarque-Bera (JB):	10.797
Skew:	-0.488	Prob(JB):	0.00452
Kurtosis:	3.359	Cond. No.	71.6

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

8.2.10. Potasio.

El potasio, al igual que en el caso de los parámetros relacionados con minerales, sufrió una reducción sustancial en la cantidad de datos; aun así, el ajuste fue aceptable, con un 73 % de la variabilidad explicada por los estimadores del índice de las bandas 1 y 11, además de la banda 5, que a su vez mostró ser estadísticamente significativos al 1 %, basados en las probabilidades de las pruebas *t*. El modelo también demostró que se ajusta a los datos por el valor del estadístico de la prueba *F*, como lo muestra la tabla (8.20). Por otra parte, la autocorrelación siguió presente en este modelo, por lo que fue necesario aplicar el modelo GLSAR.

TABLA 8.20: Resumen del modelo de regresión lineal por MCO para el potasio.

```
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.747
Model:                            OLS   Adj. R-squared:                  0.730
Method:                 Least Squares   F-statistic:                     42.91
Date:                Fri, 14 Nov 2025   Prob (F-statistic):           2.16e-09
No. Observations:                  32   AIC:                            -59.53
Df Residuals:                      29   BIC:                            -55.13
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.9446      0.074     12.690      0.000       0.792       1.097
ND_b1_b11      0.9387      0.180      5.220      0.000       0.571       1.307
b5            16.6580      2.169      7.681      0.000      12.222      21.094
==============================================================================
Omnibus:                        3.941   Durbin-Watson:                   0.977
Prob(Omnibus):                  0.139   Jarque-Bera (JB):                1.748
Skew:                          -0.198   Prob(JB):                        0.417
Kurtosis:                       1.926   Cond. No.                         144.
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Los resultados con el modelo GLSAR no fueron satisfactorios, como se evidencia en la tabla (8.21). El ajuste general del modelo fue pobre, con un R^2 ajustado de tan solo 0,266. El modelo tampoco fue estadísticamente significativo, al obtenerse una probabilidad de 0,00503 en la prueba *F*, por lo que se considera que no se ajusta a los datos. Si bien los estimadores permanecieron significativos, se vieron afectados en sus desviaciones estándar, pasando a ser significativos a un nivel de confianza del 5 %, en lugar del 1 % en el caso del modelo con MCO.

La normalidad también se pudo validar con las pruebas de *Jarque-Bera* y *Kolmogorov-Smirnov*, cuyos resultados produjeron probabilidades de 0,4935 y 0,5393, respectivamente. Asimismo, no se halló evidencia para rechazar la homocedasticidad en la varianza de los residuos, soportada por la probabilidad de 0,6957 para la prueba de *Breusch-Pagan* y de 0,1843 para la de *Goldfeld-Quandt*. El supuesto que generó el uso del modelo GLSAR, la autocorrelación, también fue resuelto al obtenerse un valor de 1,794.

TABLA 8.21: Resumen del modelo de regresión lineal por GLSAR para el potasio.

GLSAR Regression Results			
Dep. Variable:	Y	R-squared:	0.315
Model:	GLSAR	Adj. R-squared:	0.266
Method:	Least Squares	F-statistic:	6.430
Date:	Fri, 14 Nov 2025	Prob (F-statistic):	0.00503
No. Observations:	31	AIC:	-69.71
Df Residuals:	28	BIC:	-65.41

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.1120	0.109	10.238	0.000	0.890	1.335
ND_b1_b11	0.6886	0.237	2.910	0.007	0.204	1.173
b5	8.7512	3.246	2.696	0.012	2.103	15.400

Omnibus:	3.379	Durbin-Watson:	1.794
Prob(Omnibus):	0.185	Jarque-Bera (JB):	2.423
Skew:	-0.014	Prob(JB):	0.298
Kurtosis:	4.369	Cond. No.	62.5

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

A pesar de que el modelo cumple con holgura con los supuestos clásicos de la regresión lineal, el problema de falta de ajuste y la no significancia estadística del modelo llevaron a que el potasio fuese descartado del análisis, junto con el oxígeno disuelto, el fósforo y el nitrógeno.

8.2.11. Profundidad del disco Secchi.

La profundidad del disco Secchi (SDD) se ajustó inicialmente por mínimos cuadrados ordinarios, presentando un muy buen ajuste, con un valor de R^2 ajustado de 0,931. Al igual que en el caso del pH, tanto el R^2 como el R^2 ajustado arrojaron valores muy similares; por lo tanto, la adición de nuevas variables no afecta significativamente al modelo. La probabilidad del estadístico F indicó que el modelo se ajusta a los datos. Los valores de la prueba t para la evaluación de la significancia estadística de los estimadores también sugirieron que las variables independientes seleccionadas son esenciales hasta en un nivel de confianza por debajo del 1 %. La probabilidad de la prueba de *Jarque-Bera* mostró que los residuos son normales; sin embargo, la autocorrelación positiva fue muy fuerte en ese parámetro, por lo que su uso no fue adecuado en estimaciones. Los resultados completos de la regresión se presentan en la tabla (8.22).

TABLA 8.22: Resumen del modelo de regresión lineal por MCO para la SDD.

```
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.932
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                     998.3
Date:                Tue, 09 Dec 2025   Prob (F-statistic):          4.99e-127
No. Observations:                 222   AIC:                            -272.2
Df Residuals:                     218   BIC:                            -258.6
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.1375      0.020     56.813      0.000       1.098       1.177
ND_b5_b8a      4.3600      0.095     46.030      0.000       4.173       4.547
b9            22.3357      0.940     23.760      0.000      20.483      24.188
R_b4_b12      -0.1078      0.003    -35.652      0.000      -0.114      -0.102
==============================================================================
Omnibus:                        3.873   Durbin-Watson:                   0.257
Prob(Omnibus):                  0.144   Jarque-Bera (JB):                3.947
Skew:                          -0.306   Prob(JB):                        0.139
Kurtosis:                       2.773   Cond. No.                         788.
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Los mínimos cuadrados generalizados, como se ha venido experimentando, provocaron que el ajuste disminuyera; no obstante, dada la cantidad de observaciones y el resultado obtenido, que la variabilidad de los datos se explique en un 83 % indicó que el modelo siguió siendo sólido y robusto, como se muestra en la tabla (8.23). Otros parámetros que no se vieron afectados fueron las pruebas de significancia estadística, tanto de los estimadores como del modelo, produjeron probabilidades mucho menores que el nivel de confianza del 1 %. Adicionalmente, la autocorrelación se corrigió al obtenerse un valor de 1,948 para la prueba de detección de *Durbin-Watson*. Sin embargo, nótese que la probabilidad de la prueba de *Jarque-Bera* disminuyó considerablemente; esto indicó que los residuos del modelo GLSAR no siguen una distribución normal a priori.

TABLA 8.23: Resumen del modelo de regresión lineal por GLSAR para la SDD.

GLSAR Regression Results

Dep. Variable:	Y	R-squared:	0.831
Model:	GLSAR	Adj. R-squared:	0.829
Method:	Least Squares	F-statistic:	89.92
Date:	Tue, 09 Dec 2025	Prob (F-statistic):	7.53e-38
No. Observations:	221	AIC:	-640.8
Df Residuals:	217	BIC:	-627.2

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.3171	0.072	18.252	0.000	1.175	1.459
ND_b5_b8a	3.8406	0.329	11.667	0.000	3.192	4.489
b9	18.2842	3.424	5.340	0.000	11.536	25.033
R_b4_b12	-0.1159	0.007	-16.181	0.000	-0.130	-0.102

Omnibus:	91.157	Durbin-Watson:	1.948
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3170.081
Skew:	-0.842	Prob(JB):	0.00
Kurtosis:	21.478	Cond. No.	535.

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Se aplicaron las pruebas de *Shapiro-Wilk* y *Kolmogorov-Smirnov* para confirmar la no normalidad en los residuos, las cuales generaron p-valores de 0,000 y 0,0126, respectivamente. En ese orden de ideas, hay evidencia suficiente para rechazar el supuesto; no obstante, el incumplimiento de la normalidad en los residuos de los modelos realizados con mínimos cuadrados generalizados y ajustados con la técnica HAC para las desviaciones estándar no es grave en la medida en que la técnica de covarianza mencionada anteriormente asegura que tanto las probabilidades de las pruebas t como los intervalos de confianza asociados sean más robustos y confiables (Andrews, 1991).

El verdadero problema surgió con el supuesto de no heterocedasticidad. La prueba de *Goldfeld-Quandt* produjo una probabilidad de 1, lo cual fue sospechoso, debido a que las pruebas de *White* y *Breusch-Pagan* arrojaron probabilidades de 0. Esto quiere decir que la homocedasticidad no rechazada por la primera prueba infiere que la varianza no cambió abruptamente cuando se dividió la muestra; sin embargo, las dos pruebas restantes indicaron que la heterocedasticidad fue más general, incluso estructural. Esto se debe, entre otras razones, a una fuerte influencia de datos atípicos o a que las estimaciones todavía se vieron afectadas por el componente geográfico y temporal (Chen, 2016). De cualquier forma, el incumplimiento de este supuesto impidió el uso del modelo, ya que las desviaciones estándar de los parámetros y la probabilidad de la prueba F dejaron de ser confiables.

En definitiva, la SDD presentó un buen ajuste y cumplió con dos de los supuestos; no obstante, el incumplimiento de la normalidad en los residuos y de la homocedasticidad impide validar el modelo; en consecuencia, las estimaciones que se realicen no serían válidas.

8.2.12. Sodio.

Este parámetro fue modelado con un total de 33 observaciones, luego de la eliminación de datos atípicos e influyentes, generando un coeficiente de correlación ajustado de 0,826. Los índices de las bandas 2 y 7, junto con el de las bandas 5 y 8, son estadísticamente significativos a un nivel de confianza del 1 %, al igual que el modelo, dado que la probabilidad de la prueba *F* fue bastante pequeña. Los residuos de esta regresión se consideran normales gracias a la prueba de *Jarque-Bera*; sin embargo, la prueba de *Durbin-Watson* sugirió la existencia de autocorrelación positiva entre estos, como fue habitual en los modelos anteriores. El resumen completo de la regresión por mínimos cuadrados ordinarios se presenta en la tabla (8.24).

TABLA 8.24: Resumen del modelo de regresión lineal por MCO para el sodio.

```
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.837
Model:                            OLS   Adj. R-squared:                  0.826
Method:                 Least Squares   F-statistic:                     76.96
Date:                Sun, 16 Nov 2025   Prob (F-statistic):           1.54e-12
No. Observations:                  33   AIC:                            -34.34
Df Residuals:                      30   BIC:                            -29.85
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.2263      0.036    117.211      0.000       4.153       4.300
ND_b2_b7      -0.4508      0.115     -3.911      0.000      -0.686      -0.215
ND_b5_b8       1.9091      0.167     11.422      0.000       1.568       2.250
==============================================================================
Omnibus:                        5.786   Durbin-Watson:                   0.905
Prob(Omnibus):                  0.055   Jarque-Bera (JB):                2.168
Skew:                           0.239   Prob(JB):                        0.338
Kurtosis:                       1.839   Cond. No.                         8.21
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

En la tabla (8.25) se presentan los resultados por mínimos cuadrados generalizados. La autocorrelación se corrigió gracias al ajuste con el modelo GLSAR. El R^2 ajustado permaneció por encima del 70 %, por lo que se consideró aceptable para una muestra de 32 observaciones. La significancia estadística, tanto del modelo como de los estimadores, no se vio afectada para un nivel de confianza del 1 %. Asimismo, no se halló evidencia para rechazar la homocedasticidad en la varianza de los residuos, debido a que las pruebas de *Breusch-Pagan* y *White* produjeron *p*-valores de 0,626 y 0,2241, respectivamente. En sintonía con el análisis anterior, la normalidad también fue confirmada por las pruebas de *Kolmogorov-Smirnov* y *Jarque-Bera*, con probabilidades asociadas de 0,2924 y 0,3479.

Una vez validados los supuestos clásicos, se procedió a ejecutar la validación cruzada tipo *leave-one-out*, la cual arrojó un valor de 0,1382 para el error medio absoluto y de 0,1564, lo que indicó que el modelo presentó un buen ajuste con errores relativamente pequeños. Esto se confirma con

el coeficiente de determinación obtenido de 0,7685, lo que apuntó a que el modelo tiene capacidad predictiva real y confiable.

TABLA 8.25: Resumen del modelo de regresión lineal por GLSAR para el sodio.

```
==============================================================================
                         GLSAR Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.737
Model:                          GLSAR   Adj. R-squared:                  0.718
Method:                 Least Squares   F-statistic:                     17.86
Date:                Sun, 16 Nov 2025   Prob (F-statistic):           8.82e-06
No. Observations:                  32   AIC:                            -47.20
Df Residuals:                      29   BIC:                            -42.81
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.2787      0.095     44.939      0.000       4.084       4.473
ND_b2_b7      -0.5138      0.088     -5.840      0.000      -0.694      -0.334
ND_b5_b8       1.6107      0.319      5.056      0.000       0.959       2.262
==============================================================================
Omnibus:                        3.189   Durbin-Watson:                   1.642
Prob(Omnibus):                  0.203   Jarque-Bera (JB):                2.178
Skew:                           0.094   Prob(JB):                        0.337
Kurtosis:                       4.264   Cond. No.                         3.01
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

8.2.13. Temperatura.

La temperatura se modeló con un total de 157 datos, como lo muestra la tabla (8.26). El ajuste fue bastante bueno, con un R^2 ajustado de 0,97. Se demostró que el modelo se ajusta a los datos con la probabilidad de la prueba F. Asimismo, las pruebas t para la significancia estadística de los estimadores fueron satisfactorias. No obstante, se identificó un problema tanto de autocorrelación como de no normalidad en los residuos, debido a que el valor de *Durbin-Watson* está muy alejado de 0, y la probabilidad de la prueba de *Jarque-Bera* es menor que 0,05.

TABLA 8.26: Resumen del modelo de regresión lineal por MCO para la temperatura.

```
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.970
Model:                            OLS   Adj. R-squared:                  0.970
Method:                 Least Squares   F-statistic:                     2516.
Date:                Tue, 09 Dec 2025   Prob (F-statistic):          2.53e-118
No. Observations:                 157   AIC:                            -774.3
Df Residuals:                     154   BIC:                            -765.1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.0106      0.005    828.653      0.000       4.001       4.020
b5             8.7543      0.134     65.544      0.000       8.490       9.018
ND_b3_b11      0.2545      0.006     40.543      0.000       0.242       0.267
==============================================================================
Omnibus:                       10.119   Durbin-Watson:                   0.462
Prob(Omnibus):                  0.006   Jarque-Bera (JB):                7.537
Skew:                           0.422   Prob(JB):                       0.0231
Kurtosis:                       2.337   Cond. No.                        94.6
==============================================================================
```

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

Los resultados por mínimos cuadrados generalizados se presentan en la tabla (8.27). Se obtuvo un coeficiente de determinación ajustado de 0,864 mediante GLSAR, por lo que la variabilidad de las observaciones está explicada adecuadamente por el índice de las bandas 3 y 11, junto con los valores de reflectancia de la banda 5. Los estimadores arrojaron probabilidades muy pequeñas en las pruebas t, por lo que no hay evidencia para rechazar la significancia estadística de los parámetros. Adicionalmente, el valor de la prueba F para el modelo fue notablemente alto; por lo tanto, el modelo siguió ajustándose a los datos. La autocorrelación fue corregida gracias al GLSAR; sin embargo, el supuesto de normalidad, según *Jarque-Bera*, no pudo ser confirmado.

La heterocedasticidad fue un problema bastante frecuente y complejo de corregir, dado que la temperatura es un parámetro que fluctúa considerablemente con respecto a las características propias de la zona de estudio y a fenómenos ajenos como el niño o la niña (Kang and Kug, 2002). Por tal motivo, fue necesario eliminar una cantidad considerable de datos para lograr el cumplimiento del supuesto en cuestión. Finalmente, la prueba de *Goldfeld-Quandt* arrojó una probabilidad de 0,3362 y

la de *Breusch-Pagan* de 0,1620, por lo que no hubo evidencia para rechazar la homocedasticidad en la varianza de los residuos.

La validación cruzada aplicada al modelo GLSAR para la temperatura generó un error medio absoluto de 0,0194 y un error medio cuadrático de 0,0245, lo que indica que el poder predictivo del modelo fue bueno y robusto ante cambios leves en los datos. Lo anterior se sustentó también con el R^2 obtenido de 0,9563. En síntesis, si bien la cantidad de datos eliminados representó una reducción sustancial de la muestra, con los datos restantes se logró ajustar un modelo que tuviera un muy buen coeficiente de determinación, y que además cumpliera con los supuestos clásicos de la regresión lineal, a excepción de la normalidad en los residuos; no obstante, como ya se explicó en el caso de la SDD, al utilizarse el método de mínimos cuadrados generalizados con errores robustos tipo HAC, no se consideró grave el incumplimiento del supuesto en mención.

TABLA 8.27: Resumen del modelo de regresión lineal por GLSAR para la temperatura.

GLSAR Regression Results

Dep. Variable:	Y	R-squared:	0.865			
Model:	GLSAR	Adj. R-squared:	0.864			
Method:	Least Squares	F-statistic:	139.5			
Date:	Tue, 09 Dec 2025	Prob (F-statistic):	3.25e-35			
No. Observations:	156	AIC:	-924.5			
Df Residuals:	153	BIC:	-915.3			

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0455	0.019	214.242	0.000	4.008	4.083
b5	7.8952	0.936	8.440	0.000	6.047	9.743
ND_b3_b11	0.2149	0.016	13.166	0.000	0.183	0.247

Omnibus:	50.664	Durbin-Watson:	2.319
Prob(Omnibus):	0.000	Jarque-Bera (JB):	261.335
Skew:	1.042	Prob(JB):	1.79e-57
Kurtosis:	8.988	Cond. No.	50.2

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

8.2.14. Sólidos suspendidos totales.

Este parámetro se ajustó con un total de 39 datos, como se presenta en la tabla (8.28). El ajuste inicial obtenido fue de 0,74 (en el R^2 ajustado), lo cual fue bastante sólido para el tamaño de la muestra. La prueba de significancia de los estimadores arrojó resultados satisfactorios, basados en las probabilidades de las pruebas t asociadas. De la misma forma, la probabilidad de la prueba F fue bastante baja, por lo que se concluyó que el modelo se ajustó a los datos. Por primera vez, el valor de la prueba de *Durbin-Watson* indica que la autocorrelación en los residuos no fue tan fuerte; por tanto, un valor de 1,439 se consideró aceptable, evitando el uso del modelo GLSAR. El supuesto de normalidad en los residuos se pudo confirmar con las probabilidades obtenidas tanto de la prueba de *Shapiro-Wilk* y *Jarque-Bera*, las cuales fueron de 0,112 y 0,368, respectivamente. El supuesto de no heterocedasticidad se confirmó con la prueba de *Breusch-Pagan*, cuyo p-valor fue de 0,2184; por lo tanto, fue posible concluir que la varianza de los residuos es constante a lo largo de la muestra.

Finalmente, se llevó a cabo la rutina de validación cruzada *leave-one-out*, la cual produjo un error medio absoluto de 0,1884 y un error medio cuadrático de 0,2243, lo que indica que el modelo tiene una fuerte capacidad predictiva y que los estimadores generaban predicciones robustas con un error pequeño. A pesar de que el R^2 de 0,6928 fue menor que el obtenido en la regresión original, no fue señal de un decaimiento en el ajuste, pues fue un comportamiento normal en este tipo de procedimientos con modelos ajustados por MCO (Lumumba et al., 2024); por el contrario, fue una buena señal, ya que se encontró muy cerca del valor de referencia.

TABLA 8.28: Resumen del modelo de regresión lineal por MCO para los sólidos suspendidos totales.

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.754
Model:	OLS	Adj. R-squared:	0.740
Method:	Least Squares	F-statistic:	79.25
Date:	Wed, 10 Dec 2025	Prob (F-statistic):	6.50e-14
No. Observations:	39	AIC:	-8.597
Df Residuals:	36	BIC:	-3.606

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.1284	0.128	24.462	0.000	2.878	3.379
ND_b7_b8	3.2790	0.265	12.373	0.000	2.760	3.798
R_b2_b8	-0.4841	0.064	-7.513	0.000	-0.610	-0.358

Omnibus:	4.321	Durbin-Watson:	1.439
Prob(Omnibus):	0.115	Jarque-Bera (JB):	1.997
Skew:	-0.222	Prob(JB):	0.368
Kurtosis:	1.985	Cond. No.	26.0

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

8.2.15. Turbidez.

La turbidez presentó un ajuste inicial de 0,746, según el coeficiente de determinación ajustado para una muestra de 38 observaciones. El modelo se ajustó a los datos, debido a que la probabilidad de la prueba F es menor que 0,01. Los estimadores elegidos para este parámetro fueron el índice de las bandas 4 y 8, y la banda 9, los cuales presentaron evidencia de ser estadísticamente significativos, basados en las probabilidades de las pruebas t. La autocorrelación positiva en los residuos fue fuerte, según el valor de 0,534 de la prueba de *Durbin-Watson*. Adicionalmente, se aplicó la transformación logarítmica a la variable independiente, lo cual contribuyó a mejorar el ajuste y resolver problemas de heterocedasticidad. Los resultados de la regresión por mínimos cuadrados ordinarios se muestran en la tabla (8.29).

TABLA 8.29: Resumen del modelo de regresión lineal por MCO para la turbidez.

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.760
Model:	OLS	Adj. R-squared:	0.746
Method:	Least Squares	F-statistic:	55.29
Date:	Wed, 10 Dec 2025	Prob (F-statistic):	1.47e-11
No. Observations:	38	AIC:	90.63
Df Residuals:	35	BIC:	95.54

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.8508	0.155	24.876	0.000	3.537	4.165
ND_b4_b8	3.5871	0.729	4.922	0.000	2.108	5.066
b9	175.5929	27.066	6.488	0.000	120.647	230.539

Omnibus:	9.141	Durbin-Watson:	0.534
Prob(Omnibus):	0.010	Jarque-Bera (JB):	2.688
Skew:	0.209	Prob(JB):	0.261
Kurtosis:	1.766	Cond. No.	219.

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

El ajuste del modelo por mínimos cuadrados generalizados no representó una caída sustancial del R^2 ajustado, como se había evidenciado en los modelos anteriores. La significancia estadística, tanto del modelo como de los estimadores, no se afectó, debido a que las probabilidades de la prueba F y t asociadas arrojaron valores muy pequeños. La autocorrelación fue corregida exitosamente con el modelo GLSAR, con un valor de 1,715 en la prueba de *Durbin-Watson*. El resumen de la regresión se muestra en la tabla (8.30).

La varianza de los residuos se consideró homocedástica, gracias a que las pruebas de *Breusch-Pagan*, *White* y *Goldfeld-Quandt* produjeron p-valores mucho mayores que 5 %. De la misma forma, no se halló evidencia para rechazar la normalidad, debido a que las pruebas de *Kolmogorov-Smirnov* y *Jarque-Bera* arrojaron probabilidades de 0,2064 y 0,2716, respectivamente. Los resultados de la validación cruzada fueron igualmente satisfactorios. El error medio absoluto y el error medio cuadrático

fueron de 0,0194 y 0,0245, lo que indicó que el modelo era capaz de generar predicciones confiables al obtenerse errores tan pequeños. Adicionalmente, se obtuvo un coeficiente de determinación de 0,9563, por lo que la robustez y consistencia del modelo se consideraron fuertes ante pequeñas alteraciones en las observaciones.

En síntesis, el cumplimiento de los supuestos clásicos de la regresión lineal, junto con el ajuste obtenido, permitieron concluir que el modelo era apto para generar estimaciones fidedignas.

TABLA 8.30: Resumen del modelo de regresión lineal por GLSAR para la turbidez.

GLSAR Regression Results

Dep. Variable:	Y	R-squared:	0.757
Model:	GLSAR	Adj. R-squared:	0.743
Method:	Least Squares	F-statistic:	27.58
Date:	Wed, 10 Dec 2025	Prob (F-statistic):	7.62e-08
No. Observations:	37	AIC:	57.95
Df Residuals:	34	BIC:	62.78

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0729	0.363	11.217	0.000	3.335	4.811
ND_b4_b8	3.4038	0.876	3.887	0.000	1.624	5.183
b9	143.9881	22.903	6.287	0.000	97.443	190.533

Omnibus:	2.358	Durbin-Watson:	1.715
Prob(Omnibus):	0.308	Jarque-Bera (JB):	1.373
Skew:	0.008	Prob(JB):	0.503
Kurtosis:	3.944	Cond. No.	70.4

Fuente: Elaboración propia mediante regresión lineal en Python utilizando *statsmodels*.

8.3. Resumen de los modelos finales.

Una vez ajustados y validados todos los modelos, se pudo concluir que los componentes espacio-temporales tuvieron una incidencia directa sobre los resultados y, más específicamente, sobre los residuos. De igual manera, el ajuste por mínimos cuadrados generalizados con autocorrelación demostró ser una herramienta muy poderosa y confiable al momento de neutralizar dicha incidencia, ayudado por la aplicación de errores robustos con la técnica HAC. A pesar de que 10 parámetros pudieron ser modelados exitosamente, hubo otro grupo de 5 que presentaron problemas que no pudieron ser resueltos mediante técnicas de regresión lineal. Lo anterior no sugiere que dichas características no pudieran ser modeladas, sino que su comportamiento en la zona de estudio no permitió el uso de estructuras lineales como enfoque principal para diagramar su conformación original.

La tabla (8.31) presenta los modelos finales que cumplieron con los supuestos clásicos de la regresión lineal y que generaron un ajuste aceptable, basado en el coeficiente de determinación ajustado. El parámetro con la mayor correlación fue la temperatura, cuyos estimadores explican en un 86,4 % la variabilidad del logaritmo de la variable independiente, cabe aclarar que este parámetro, junto con el pH, se ajustó con una gran cantidad de datos, por lo que era de esperarse que tuviera un buen desempeño en la regresión. El parámetro con el coeficiente más bajo fue el sodio, en el cual la variabilidad del logaritmo de la variable respuesta fue explicada en un 71,8 %.

TABLA 8.31: Resumen de los modelos significativos.

Parámetro	**Modelo**	R^2
Arsénico	$y^3 = 0,3069 + 0,0062 * \frac{b8a - b9}{b8a + b9} + 0,2488 * \frac{b1 - b5}{b1 + b5} - 0,0135 * \frac{b9 - b12}{b9 + b12}$	0,747
Sulfato	$Ln(y) = 3,0966 + 2,8283 * \frac{b5 - b11}{b5 + b11} - 2,3679 * \frac{b8 - b11}{b8 + b11}$	0,753
Clorofila-a	$Ln(y) = 2,952 - 2,6178 * \frac{b2 - b7}{b2 + b7} + 2,9818 * \frac{b3 - b8}{b3 + b8}$	0,745
Cloruro	$Ln(y) = 4,5992 + 2,2868 * \frac{b5 - b11}{b5 + b11} - 1,8139 * \frac{b8 - b11}{b8 + b11}$	0,747
Dureza	$y = 97,7492 + 217,5363 * \frac{b5 - b11}{b5 + b11} - 178,3482 * \frac{b8 - b11}{b8 + b11}$	0,731
pH	$y = 7,0804 + 1,0089 * \frac{b5 - b8a}{b5 + b8a} + 22,1631 * b1$	0,746
Sodio	$Ln(y) = 4,3062 - 0,5211 * \frac{b2 - b7}{b2 + b7} + 1,7034 * \frac{b5 - b8}{b5 + b8}$	0,718
Temperatura	$Ln(y) = 4,0455 + 0,2149 * \frac{b3 - b11}{b3 + b11} + 7,8952 * b5$	0,864
TSS	$Ln(y) = 3,1206 + 3,2588 * \frac{b3 - b11}{b3 + b11} - 0,4841 * \frac{b2}{b8}$	0,74
Turbidez	$y = 4,0729 + 3,4038 * \frac{b4 - b8}{b4 + b8} + 143,9883 * b9$	0,743

Fuente: Elaboración propia.

En términos generales, la mayoría de los modelos permanecieron alrededor del 74 % de variabilidad explicada, y teniendo en cuenta la cantidad de datos con la que se realizó el ajuste, se pudo concluir que la respuesta espectral de las muestras no estuvo estrechamente relacionada con su comportamiento físico-químico inherente, debido a la gran cantidad de observaciones que fue necesario remover para obtener dicho ajuste. Adicionalmente, el obstáculo de la autocorrelación sugirió que

un enfoque lineal era equívoco en la medida en que solamente un modelo pudo ser modelado por medio de mínimos cuadrados ordinarios; no obstante, este podría ser un punto de partida para la exploración y el uso de alternativas, como modelos ponderados (o con pesos) o series de tiempo, de tal manera que se tome en cuenta el componente espacio-temporal que imposibilitó que varios parámetros mostraran una correlación decente.

Como se puede apreciar en la tabla (8.32), la razón que predomina en la decisión de descartar los modelos estuvo fuertemente relacionada con la autocorrelación en los residuos de las estimaciones, dado que tanto el fósforo como el nitrógeno y el potasio presentaron coeficientes de determinación ajustados muy bajos al utilizar los mínimos cuadrados generalizados, en contraste con sus similares, como el sulfato o el arsénico, cuyo R^2 disminuyó de igual manera, pero no hasta el punto de entorpecer el ajuste general del modelo. El oxígeno disuelto, por otra parte, logró un gran ajuste junto con el cumplimiento de los supuestos clásicos; no obstante, se decidió apartarlo del análisis debido a que dicho ajuste se logró luego de la eliminación de más del 75 % de la muestra, lo que indicó que el comportamiento espectral de las observaciones no siguió una estructura lineal modelable. Finalmente, la varianza de los residuos del modelo de la profundidad del disco Secchi mostró evidencia de tener una heterocedasticidad estructural fuerte, lo que significó que el error varía a medida que se recorre la muestra; por lo tanto, no pudo ser incluido debido a la incertidumbre tanto en el ajuste como en las estimaciones que genera el incumplimiento de este supuesto.

TABLA 8.32: Resumen de los modelos no significativos.

Parámetro	**Ajuste**	**Razón**
Oxígeno disuelto	0,729	Datos eliminados y no normalidad en los residuos
Fósforo	0,179	Bajo ajuste y problemas de autocorrelación
Nitrógeno	-0,021	Bajo ajuste y problemas de autocorrelación
Potasio	0,266	Bajo ajuste y problemas de autocorrelación
SDD	0,829	Problemas de heterocedasticidad estructural

Fuente: Elaboración propia.

9

Predicción de Parámetros.

El propósito inicial del ajuste de los modelos de regresión lineal estaba relacionado con la estimación de los parámetros de calidad del agua en el lago George; por consiguiente, fue necesario determinar los puntos de análisis en los cuales fuera posible extraer la información espectral que serviría de insumo para su evaluación en las ecuaciones de las regresiones. Por lo tanto, se generó una muestra aleatoria de 100.000 puntos en la zona de estudio, con la restricción de que solamente podía haber un punto por píxel, de tal forma que se garantizara extraer solamente un valor de reflectancia, evitando redundancias en las predicciones. La muestra generada se presenta en la figura (9.1).

Cabe resaltar que el uso de un modelo de regresión lineal, de cualquier tipo, impone restricciones para garantizar la validez de los resultados. Una de las más relevantes es asegurar que los valores de las variables se encuentren dentro del rango con el que inicialmente se generaron los modelos; es decir, si una regresión se desarrolló con una variable predictora x cuyos valores van del 0 al 1, las estimaciones que se vayan a obtener no deben generarse con valores de x fuera de ese rango. De lo contrario, se estaría extrapolando predicciones a partir de una ecuación que no fue entrenada con valores de dicha clase. En consecuencia, la muestra de los puntos sobre el lago debe restringirse para cada modelo, índice, cociente o banda utilizada en cada parámetro.

9.1. Validación de Rangos y Extracción de Valores de Reflectancia.

El procedimiento para extraer los valores de reflectancia que fueran válidos para ser evaluados en los modelos siguió la misma estructura para todos los parámetros. En primer lugar, se determinó el rango de las variables de cada ecuación; para el arsénico, por ejemplo, el índice entre las bandas 1 y 5 fue de -0,4091 a 0,4667. Luego, se leyeron los valores de reflectancia solamente de las bandas que participaron, que para el caso del sulfato fueron las bandas 5, 8 y 11. Se calcularon los índices correspondientes y se validó que estuvieran dentro del intervalo correspondiente. Una vez que el punto cumplía con el rango para las 2 o 3 variables asociadas, se guardaba la información extraída en una base de datos temporal. El procedimiento anteriormente explicado se repitió para las 20 imágenes y para los 10 modelos que fueron verificados. Los rangos para todas las variables de todos los parámetros se muestran en la tabla (9.1).

TABLA 9.1: Rango de las variables de cada modelo.

Parámetro	Variable	Rango
Arsénico	ND_b8a_b9	-1,5106 a 63,00
	ND_b1_b5	-0,4091 a 0,4667
	ND_b9_b12	-22,00 a 8,33

Parámetro	Variable	Rango
Sulfato	ND_b5_b11	0,3109 a 0,9286
	ND_b8_b11	0,2026 a 0,9333
Clorofila-a	ND_b2_b7	-0,7647 a 0,6170
	ND_b3_b8	-1,0714 a 0,5330
Cloruro	ND_b5_b11	0,3109 a 0,9286
	ND_b8_b11	0,1496 a 0,9333
Dureza	ND_b5_b11	0,3109 a 0,9286
	ND_b8_b11	0,1496 a 0,9333
pH	ND_b5_b8a	0,0000 a 0,5820
	b1	0,0015 a 0,0396
Sodio	ND_b2_b7	-0,7647 a 0,4039
	ND_b5_b8	-0,1837 a 0,5691
Temperatura	ND_b3_b11	0.000 a 0,9155
	b5	0,002 a 0,0393
TSS	ND_b7_b8	-0,3182 a 0,2424
	R_b5_b8	0,069 a 3,549
Turbidez	ND_b4_b8	-0,3636 a 0,4421
	b9	-0,0068 a 0,0143

Fuente: Elaboración propia.

Dado que un número considerable de variables presentaba rangos bastante limitados, la cantidad de puntos en imágenes específicas era notablemente baja; por ejemplo, el modelo de turbidez solamente evidenció 389 puntos válidos de 100.000, según los rangos de sus variables independientes para el mes de junio. Otro parámetro que generó problemas en la cantidad tan dispersa de puntos evaluables fue el sulfato, con 21.916 ubicaciones válidas, luego de eliminar casi la mitad de las imágenes, como se muestra en la tabla (9.2). Nótese cómo para el cloruro y la dureza se encontró la misma cantidad de puntos; esto se debe a que el rango de los modelos de ambas variables dependientes era el mismo, lo que supone que ambos solo pueden ser evaluados en las mismas secciones.

Teniendo en cuenta que el índice de calidad del agua se calcula por ubicación y no por parámetro, se descartaron aquellas fechas (imágenes) con la menor cantidad de puntos encontrados, siendo el sulfato, el cloruro y la dureza los minerales con la mayor cantidad de imágenes desestimadas. La clorofila-a, por otra parte, demostró ser bastante consistente a lo largo de la muestra espectral, con tan solo una imagen removida. Otros parámetros que resultaron ser bastante sólidos a lo largo del periodo del análisis fueron el sodio y los sólidos suspendidos totales, a los cuales se les redujo la muestra en tan solo dos épocas.

TABLA 9.2: Puntos válidos por parámetro.

Parámetro	Puntos válidos	Épocas removidas
Arsénico	37.962	5
Sulfato	21.916	9

Parámetro	Puntos válidos	Épocas removidas
Clorofila-a	63.549	1
Cloruro	17.514	8
Dureza	17.514	8
pH	27.372	7
Sodio	37.525	2
Temperatura	31.067	6
TSS	34.065	2
Turbidez	26.549	7

Fuente: Elaboración propia.

Hasta este punto, se habían obtenido las localizaciones donde cada modelo independiente es evaluable; sin embargo, fue imperativo cotejar cada base de datos temporal para determinar la cantidad de puntos en los que todos los parámetros pudieran ser estimados. De esta forma, sería posible calcular el índice de calidad del agua con la totalidad de los modelos disponibles. Mediante un proceso de filtrado básico, se halló que un total de 702 puntos comparten ubicación dentro de la muestra, lo que representa un 0,7 % de la cantidad global utilizada en el análisis inicial.

Como se puede observar en la figura (9.2), los puntos evaluables están distribuidos principalmente en la zona oriental del lago; no obstante, no se encuentran agrupados en un solo sector, sino que abarcan un área considerable de norte a sur. El hecho de que una parte significativa del lago no estuviera contemplada por las ubicaciones evaluables no fue alarmante en el sentido de que dichas zonas podrían ser utilizadas como parámetro de evaluación en el modelo que posteriormente sería ajustado con los valores del índice de calidad del agua calculado.

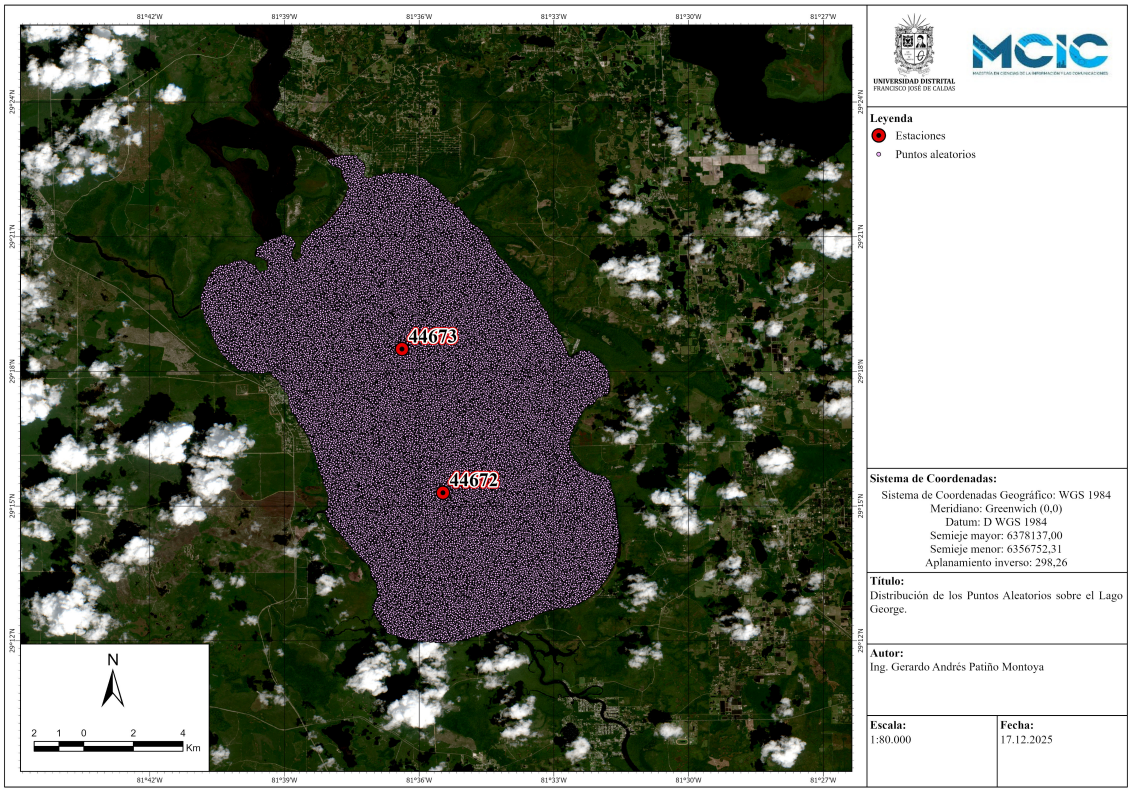


FIGURA 9.1: Ubicación de las estaciones en el Lago George.
Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023).

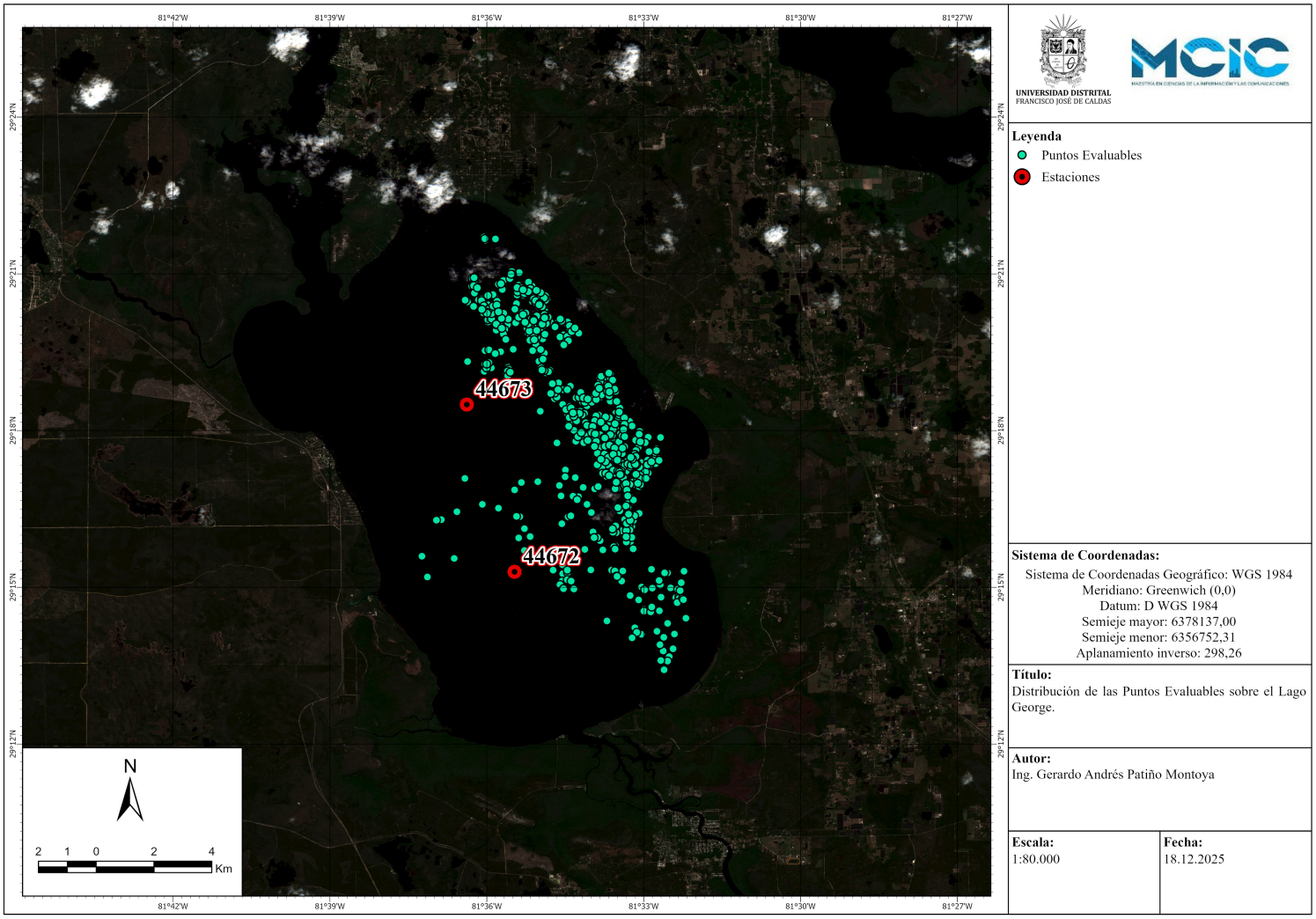


FIGURA 9.2: Distribución de los evaluables sobre el lago George.
Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023).

Finalmente, los valores de reflectancia de las bandas correspondientes fueron extraídos para los 702 puntos evaluables que se filtraron para cada parámetro independiente. Asimismo, se calcularon los índices y cocientes necesarios según la estructura de cada regresión. La tabla (9.3) muestra la cantidad de muestras finales que posteriormente serían objeto de estimación por variable; este número es el resultado de multiplicar las ubicaciones mencionadas anteriormente por el número de imágenes, después del proceso de filtrado; por consiguiente, la clorofila fue el parámetro con la mayor cantidad de predicciones, con 13.338, y el sulfato con la menor, con tan solo 7.722.

TABLA 9.3: Cantidad de muestras finales.

Parámetro	**Muestra total**
Arsénico	10.530
Sulfato	7.722
Clorofila-a	13.338
Cloruro	8.424
Dureza	8.424
pH	9.126
Sodio	12.636
Temperatura	9.828
TSS	12.636
Turbidez	9.126

Fuente: Elaboración propia.

9.2. Predicciones.

Una vez se consolidaron las épocas y los puntos cotejados, se procedió a evaluar cada modelo. La tabla (9.4) muestra el resumen de las estadísticas básicas de las estimaciones. El valor de la izquierda en la columna de mínimos, máximos, rangos, desviaciones estándar y promedios representa el cálculo de los resultados del modelo, mientras que los valores de la derecha de las columnas anteriormente mencionadas simbolizan las estadísticas de la base de datos original con la que se realizó el ajuste de cada regresión.

Los resultados del arsénico fueron bastante consistentes, tanto el mínimo como el máximo y el rango estuvieron notablemente cerca de los valores de referencia; adicionalmente, la desviación estándar sugirió que las estimaciones son menos dispersas que las muestras originales. Para el sulfato se obtuvieron predicciones mayores al límite de la muestra inicial; no obstante, la dispersión disminuyó casi a la mitad, así como el promedio, que decreció en 4 unidades. La clorofila mostró ser consistente en valores bajos al estimar concentraciones ligeramente por debajo del límite de referencia; sin embargo, la combinación de variables para este modelo hizo que se presentaran estimaciones muy por encima del valor máximo de la muestra principal; aun así, la desviación estándar y el promedio estuvieron cerca de los cálculos originales.

Asimismo, el cloruro y la dureza tuvieron resultados similares, lo cual era de esperarse debido a que ambos modelos se ajustaron con las mismas variables independientes; por parte del cloruro, todas las medidas presentaron consistencia al estar bastante cerca de los valores de referencia salvo por ciertos valores máximos que superaron el umbral de la muestra inicial. En ambos casos, la desviación estándar se redujo significativamente, lo que demuestra la estabilidad de las estimaciones. Otro

parámetro que mostró una solidez ineludible fue la acidez mediante el pH, cuyas predicciones estuvieron dentro del rango de referencia y tanto la desviación estándar como el promedio disminuyeron ligeramente en comparación con la muestra originaria.

Por último, los resultados de los modelos del sodio, la temperatura y la turbidez demostraron una consistencia similar al comparar los respectivos valores de referencia, presentando rangos, desviaciones estándar y promedios parecidos, o incluso, menores a los de la muestra original. No obstante, las estimaciones de los sólidos suspendidos totales tuvieron un comportamiento similar al de la clorofila en la medida que las predicciones mínimas no sobrepasaron el umbral, pero las máximas sí, y con creces; mientras la medición máxima de los datos iniciales fue de 20,6, para las estimaciones fue de 37,45, lo que genera que el rango se vea afectado considerablemente. Aun así, la desviación estándar del modelo se redujo casi a la mitad, con el promedio levemente incrementado sin señales alarmantes de ajuste erróneo.

TABLA 9.4: Cantidad de muestras finales.

Parámetro	Mínimos	Máximos	Rangos	Desv. Estándar	Promedios
Arsénico	0,55 - 0,52	0,92 - 0,92	0,38 - 0,4	0,03 - 0,1	0,69 - 0,71
Sulfato	18,16 - 16,03	74,84 - 66,94	56,68 - 50,91	8,8 - 15,53	38,07 - 43,39
Clorofila-a	2,4 - 3,73	306,95 - 48,46	304,55 - 44,73	10,5 - 11,17	21,02 - 22,69
Cloruro	88,95 - 77,34	278,42 - 264,79	189,47 - 187,45	29,36 - 52,51	159,8 - 176,47
Dureza	84,39 - 69,98	193,16 - 208,86	108,77 - 138,88	17,23 - 39,07	138,61 - 151,56
pH	7,17 - 7,16	8,43 - 8,68	1,27 - 1,52	0,26 - 0,44	7,75 - 7,93
Sodio	45,25 - 40,51	151,26 - 166,71	106,01 - 126,2	18,12 - 30,1	91,19 - 99,51
Temperatura	58,89 - 56,12	82,06 - 87,14	23,16 - 31,02	3,98 - 8,54	69,9 - 74,35
TSS	4,91 - 3,6	37,45 - 20,6	32,54 - 16,99	3,34 - 4,19	13,99 - 11,33
Turbidez	2,38 - 1,43	7,12 - 7,33	4,75 - 5,9	0,78 - 1,52	4,78 - 4,38

Fuente: Elaboración propia.

En síntesis, salvo por los valores particularmente altos estimados por los modelos de los sólidos suspendidos totales y la clorofila-a, los resultados fueron bastante consistentes y en línea con las bases de datos relacionadas con las mediciones sobre el lago George para el periodo en análisis para las 702 ubicaciones filtradas. Con base en los productos anteriormente explicados, se procedió a hacer el cálculo del índice de calidad del agua del Consejo Canadiense de Ministros del Medio Ambiente, desarrollado en la siguiente sección.

10

Cálculo del Índice de Calidad del Agua.

El cálculo del índice de calidad del agua, siguiendo la metodología del CCME, se completó en dos fases principales. En la primera, se obtuvo la cantidad de estimaciones que estuvieron por encima (si el objetivo era máximo) o por debajo (si el objetivo era mínimo) del umbral dado para un parámetro específico; de esta manera, se determinaba el alcance, con la cantidad de muestras que fallaron; la frecuencia, con la cantidad de parámetros que tuvieron al menos una observación no conforme; y la amplitud, con la diferencia ponderada entre los valores objetivos y la magnitud de las pruebas reprobadas. Luego de evaluar todos los parámetros, se calcularon todos los factores globales, se ponderaron los valores del índice y se categorizó para cada ubicación.

10.1. Fase 1: Cálculo de los Factores por Parámetro.

Como ya se mencionó anteriormente, el cálculo del índice depende de tres factores fundamentales: la frecuencia, el alcance y la amplitud. Cada uno de estos factores depende directamente de la cantidad y la magnitud de las observaciones que presentaron estimaciones fuera del rango considerado válido para cada variable. A continuación, se sintetizan los resultados del análisis, las implicaciones y las posibles consecuencias de las concentraciones excesivas en el lago George; asimismo, se presenta la fuente de los valores objetivo para cada indicador con su respectiva justificación.

10.1.1. Arsénico.

El arsénico es un mineral que no presentó mediciones considerablemente altas en el lago George. Esta misma tendencia se observó en la determinación de las pruebas fallidas, puesto que no solamente ninguna prueba falló, sino que la observación más alta fue diez veces menor que el valor objetivo. Esto quiere decir que el cuerpo lacustre objeto de análisis no tiene una mineralización arsenical significativa, la cual es inducida principalmente por la actividad agrícola e industrial (Tsai et al., 1999). En consecuencia, este mineral no tuvo ningún impacto sobre el índice, debido a que ni el alcance, ni la frecuencia, ni la amplitud penalizaron al no haber ninguna observación por encima del umbral, como se puede observar en la tabla (10.1).

TABLA 10.1: Resumen de las pruebas fallidas del arsénico.

Medida	**Valor**
Unidades	$\mu g/L$
Objetivo	10,0
Tipo	Máximo
Mínimo	0,548

Medida	Valor
Máximo	0,923
Muestra	10.530
Obs. Fallidas	0,0
Porcentaje	0,0

Fuente: Elaboración propia.

El valor objetivo del arsénico se obtuvo directamente de la Agencia de Protección Ambiental de Estados Unidos, la cual indica que los efectos potenciales ante una exposición prolongada en ambientes con concentraciones muy altas de este mineral pueden representar problemas en la piel y en el sistema circulatorio. Asimismo, incrementa la probabilidad de desarrollar cáncer. Las fuentes contaminantes, como ya se había mencionado anteriormente, están relacionadas con el desprendimiento de depósitos naturales por erosión y con los desechos industriales en la fabricación de vidrio o productos electrónicos (EPA, 2006).

10.1.2. Sulfato.

Las concentraciones de sulfato en el lago tampoco fueron significativamente altas, incluso teniendo una variabilidad notable en las estimaciones, con un rango de aproximadamente 56 mg/L. Esto significa que la cantidad del mineral fluctúa naturalmente, pero no incide directamente en la calidad del agua del cuerpo. En un caso similar al del arsénico, el valor máximo estimado no se acercó al umbral determinado, por lo que se consideró que el parámetro tampoco afectó al índice de alguna forma. En síntesis, tanto la amplitud como la frecuencia y el alcance fueron nulos. Los resultados del análisis se presentan en la tabla (10.2).

TABLA 10.2: Resumen de las pruebas fallidas del sulfato.

Medida	Valor
Unidades	mg/L
Objetivo	250,0
Tipo	Máximo
Mínimo	18,161
Máximo	74,839
Muestra	7.722
Obs. Fallidas	0,0
Porcentaje	0,0

Fuente: Elaboración propia.

El valor objetivo del sulfato es de 250 mg/L, el cual fue adoptado por los lineamientos de la Agencia de Protección Ambiental de los Estados Unidos, que considera al sulfato o sulfato como un estándar secundario. Estos estándares están enfocados en guiar a las organizaciones de potabilización de agua en la eliminación de químicos en el agua que puedan afectar tanto el color como el sabor, e incluso el olor, del líquido que se está consumiendo. Para el caso del parámetro en cuestión, una alta concentración significaría un cambio en el sabor, tendiendo a ser más salado (EPA, 2023).

10.1.3. Clorofila-a.

Concretamente, el parámetro que afectó en mayor medida el índice de calidad del agua en el lago fue el de la clorofila-a. De un total de 13.338 predicciones, 10.043 fallaron la prueba, representando un 75,3 % de la muestra. Si bien el modelo presentó valores abrumadoramente altos (mayores que 300), el promedio general de las estimaciones estuvo ligeramente por debajo de los números con los que originalmente se ajustó la regresión; asimismo, la desviación estándar fue inferior; por lo tanto, los valores mencionados correspondieron a predicciones atípicas que representaban una medición extrema específica, codificada por la combinación de bandas. En otras palabras, el hecho de tener estimaciones altas no significó que el modelo presentara problemas de evaluación, sino que las características propias del cuerpo de agua indujeron la existencia de un fenómeno eutrófico mayor que se tradujo en la obtención de observaciones excesivas.

En cuanto al cálculo de los factores, el parámetro ya comenzaba a afectar al alcance por tener al menos una prueba fallida, lo cual cumplió con creces. La frecuencia se vio terriblemente degradada por la notable proporción de pruebas que superaron el valor objetivo; por tanto, el factor se disparó, arrastrando el índice hacia valores bajos. La amplitud es el factor más perjudicado por los valores excesivos anteriormente mencionados, lo que condujo a la obtención de excursiones sustancialmente enormes. El resumen de los resultados se muestra en la tabla (10.3).

TABLA 10.3: Resumen de las pruebas fallidas de la clorofila-a.

Medida	**Valor**
Unidades	$\mu g/L$
Objetivo	15,0
Tipo	Máximo
Mínimo	2,395
Máximo	306,946
Muestra	13.338
Obs. Fallidas	10.043
Porcentaje	75,3

Fuente: Elaboración propia.

La Agencia de Protección Ambiental de los Estados Unidos ha publicado una serie de reportes técnicos en los que evalúa el comportamiento ambiental, ecológico y biológico en los cuerpos de agua lacustres desde el año 2007 hasta 2022. En el (EPA, 2012) se determinó que las concentraciones superiores a 15 $\mu g/L$ pueden conducir al desarrollo de eutrofia fuerte, lo que se traduce en el crecimiento de plantas y la proliferación masiva de algas, incidiendo directamente en la calidad del agua.

10.1.4. Cloruro.

Para este compuesto, se obtuvieron estimaciones más cercanas al valor objetivo en comparación con los parámetros anteriores; incluso, hubo un total de 25 de 8.424 observaciones que lo superaron, como se muestra en la tabla (10.4). Aun así, las pruebas fallidas solo representaron un 0,3 % del total de la muestra. En términos prácticos, el alcance se vio afectado desde el principio, puesto que el parámetro global falló; sin embargo, la frecuencia mencionada anteriormente no castigó tajantemente el índice. Adicionalmente, la predicción máxima de 278,42 mg/L tampoco representa una afectación sustancial, dado que las excursiones serían considerablemente pequeñas.

TABLA 10.4: Resumen de las pruebas fallidas del cloruro.

Medida	Valor
Unidades	mg/L
Objetivo	250,0
Tipo	Máximo
Mínimo	88,951
Máximo	278,417
Muestra	8.424
Obs. Fallidas	25
Porcentaje	0,3

Fuente: Elaboración propia.

El umbral elegido para este parámetro fue de 250 mg/L, según el listado de estándares secundarios de la Agencia de Protección Ambiental de los Estados Unidos (EPA, 2023), la cual indica que las altas concentraciones de cloruro en un cuerpo de agua lacustre están relacionadas con acumulaciones salinas focalizadas por temporadas, posiblemente provenientes de aportes subterráneos mineralizados naturales o por procesos de evaporación elevada. Independientemente de las causas, los valores detectados no determinan una amenaza real a la calidad del agua del lago, con base en el impacto en el índice.

10.1.5. Dureza.

El comportamiento de este parámetro resultó ser bastante similar al del cloruro en la medida en que las concentraciones estimadas estuvieron bastante cerca del umbral, presentando un total de 66 pruebas por encima de este, tal como se evidencia en la tabla (10.5). El impacto de este componente en el índice también fue bajo, debido a que el valor máximo de las predicciones estuvo a 13.15 mg/L del valor objetivo, por lo que las excursiones no penalizaron en mayor medida la amplitud. De forma congruente, solamente el 0,8 % de las observaciones superaron el límite sugerido; en consecuencia, la frecuencia tampoco se vio afectada sustancialmente. En síntesis, si bien el índice se vio perjudicado, no fue en gran medida debido a la escasa cantidad y a los valores asociados de las pruebas fallidas.

TABLA 10.5: Resumen de las pruebas fallidas del cloruro.

Medida	Valor
Unidades	mg/L
Objetivo	180,0
Tipo	Máximo
Mínimo	84,389
Máximo	193,156
Muestra	8.424
Obs. Fallidas	66
Porcentaje	0,8

Fuente: Elaboración propia.

La dureza está relacionada con las concentraciones de magnesio y calcio en el agua, los cuales son minerales saludables para el consumo humano; sin embargo, una cantidad elevada de estos elementos en el líquido puede representar un daño grave en las tuberías metálicas y en los elementos de uso

común, como electrodomésticos o redes de acueducto. Por lo tanto, el Servicio Geológico de los Estados Unidos (USGS) determinó que las acumulaciones mayores a 180 mg/L de estos dos minerales pueden afectar notablemente la vida útil de los objetos que entren en contacto con el agua (USGS, 2018).

10.1.6. Acidez (pH).

Teniendo en cuenta las mediciones recolectadas en la zona de estudio y las estimaciones generadas por el modelo de regresión lineal, fue posible concluir que los niveles de acidez, según el pH del lago, fueron bastante estables, sin signos ni evidencia de picos temporales o geográficos. Esto se vio reflejado en la evaluación de los factores, puesto que de 9.126 predicciones, ninguna superó el umbral, tanto mínimo como máximo; por lo que la frecuencia no tuvo peso alguno en el índice. Asimismo, al no haber ningún resultado incorrecto, el alcance tampoco se vio perjudicado. Finalmente, las excursiones de este parámetro no tuvieron incidencia al ser nulas; por lo tanto, la amplitud se canceló al igual que los demás factores. El resumen de los resultados se muestra en la tabla (10.6).

TABLA 10.6: Resumen de las pruebas fallidas del pH.

Medida	**Valor**
Unidades	Unidad de pH
Objetivo	6,5 - 8,5
Tipo	Mínimo y máximo
Mínimo	7,167
Máximo	8,432
Muestra	9.126
Obs. Fallidas	0,0
Porcentaje	0,0

Fuente: Elaboración propia.

La medición del pH está categorizada, según la Agencia de Protección Ambiental de los Estados Unidos, en estándares secundarios de agua potable. Un líquido con demasiada acidez (8,5 o superior en la escala de medición de pH) suele tener un sabor agrio con tintes metálicos. Además, tiene implicaciones al contacto con objetos metálicos por la corrosión que genera. Paralelamente, un líquido con valores bajos de pH (6,5 o inferior en la escala de medición de pH) presenta una consistencia resbaladiza, casi gelatinosa, con una sensación gaseosa (EPA, 2023).

10.1.7. Sodio.

El agua del lago George no se categoriza completamente dentro de los cuerpos de agua dulce, debido a que varios afluentes subterráneos provenientes del Océano Atlántico confluyen a su alrededor, contribuyendo a la acumulación de depósitos salinos, como ya se evidenció con las medidas obtenidas para el cloruro. En ese orden de ideas, era de esperarse que las concentraciones de sodio tuvieran un impacto significativo sobre los factores del índice; no obstante, la evidencia mostró que ninguna de las pruebas superó el umbral escogido para este parámetro; por lo tanto, la frecuencia, la amplitud y el alcance fueron anulados al no fallar ninguna prueba de las 12.636 realizadas. Los resultados de la evaluación del sodio se presentan en la tabla (10.7).

TABLA 10.7: Resumen de las pruebas fallidas del sodio.

Medida	Valor
Unidades	mg/L
Objetivo	200
Tipo	Máximo
Mínimo	45,247
Máximo	151,26
Muestra	12.636
Obs. Fallidas	0,0
Porcentaje	0,0

Fuente: Elaboración propia.

El valor objetivo para este parámetro fue de 200 mg/L. Este umbral no fue adoptado por la Agencia de Protección Ambiental de los Estados Unidos, debido a que dicha organización no tiene una medida estándar para el sodio, puesto que es un mineral de consumo básico y difícil de medir en la dieta del ser humano. A pesar de esto, el gobierno canadiense sí determinó que una acumulación mayor a la mencionada anteriormente puede ser perjudicial en personas con problemas de hipertensión o insuficiencia cardíaca congestiva; en consecuencia, requiere una dieta restrictiva de sodio (Government of Canada, 1992).

10.1.8. Temperatura.

El estado de Florida es bien conocido por experimentar altas temperaturas, incluso en la temporada de invierno norteamericano, gracias a su privilegiada localización, topografía principalmente plana y altitud promedio muy cercana al nivel del mar. Por tanto, se especulaba sobre la obtención de mediciones potencialmente altas en cuanto a la temperatura se refiere; sin embargo, los resultados no mostraron evidencia de que este parámetro fuera sustancialmente crítico para el agua del lago George, en la medida en que ninguna observación estimada estuvo por encima del umbral sugerido. Esto quiere decir que los factores no incidieron de ninguna forma en el índice de calidad al presentarse 0 pruebas fallidas, tal como se resume en la tabla (10.8).

TABLA 10.8: Resumen de las pruebas fallidas de la temperatura.

Medida	Valor
Unidades	$°F$
Objetivo	86
Tipo	Máximo
Mínimo	58,894
Máximo	82,056
Muestra	9.828
Obs. Fallidas	0,0
Porcentaje	0,0

Fuente: Elaboración propia.

El valor objetivo de este parámetro tiene una interpretación bastante ambigua dentro del análisis de calidad de agua, debido a que no existe un estándar oficial de ninguna institución o agencia que

permita determinar la incidencia de mediciones excesivas sobre el agua. Entre otras razones, por el hecho de que el monitoreo de la temperatura difiere y se ajusta dependiendo de las características propias del territorio en análisis; por ejemplo, los estándares de un país mayoritariamente frío como Canadá, con lagos que se congelan en invierno, no son los mismos que los de un estado como Florida, en el cual la temperatura más baja en cuerpos de agua lacustres no es inferior a 56 $°F$ (o 13 $°C$). Ante este panorama, el umbral fue adoptado de (EPA, 1986), el cual es un documento de criterios de calidad del agua en el que se analiza el impacto de la temperatura y la acidez en la sensación y sabor del líquido, concluyendo que temperaturas menores que 32 $°F$ (0 $°C$), o mayores que 86 $°F$ (30 $°C$) pueden alterar el sabor del agua de manera significativa; no obstante, no alteran sus propiedades físicas o químicas.

10.1.9. Sólidos suspendidos totales.

El análisis de los TSS cobra gran importancia tras los hallazgos de las estimaciones de la clorofila, puesto que este parámetro también está relacionado con la claridad del agua en el lago. Para un total de 12.636 pruebas, se obtuvo que 84 estuvieron por encima del umbral establecido, representando un 0,7 % de la muestra, lo que indicó que el problema de claridad en el agua no se debió a una carga sedimentaria fuerte, sino a procesos exclusivamente biológicos y eutróficos. En otras palabras, el cuerpo lacustre presentó una gran concentración de algas, pero no de sólidos. La frecuencia se vio nuevamente afectada por las pruebas fallidas; sin embargo, la amplitud no penalizó en mayor medida, ya que los valores máximos encontrados superaron el valor objetivo solamente en 8 mg/L, lo cual se tradujo en excursiones pequeñas. Los resultados del análisis para los TSS se resumen en la tabla (10.9).

TABLA 10.9: Resumen de las pruebas fallidas de los TSS.

Medida	**Valor**
Unidades	mg/L
Objetivo	25,0
Tipo	Máximo
Mínimo	4,909
Máximo	37,447
Muestra	12.636
Obs. Fallidas	84
Porcentaje	0,7

Fuente: Elaboración propia.

Los sólidos suspendidos totales no solo afectan la claridad del agua, sino también el desarrollo de la vida acuática en los cuerpos hídricos que presentan acumulaciones excesivamente altas. La Agencia de Protección Ambiental de los Estados Unidos, en el reporte de Criterios de la Calidad del Agua de 1976, evaluó el impacto de valores de TSS superiores a 80 mg/L, evidenciando una reducción del 60 % en la población de microorganismos invertebrados. En consecuencia, se determinó que, hasta los 25 mg/L, las concentraciones no afectan de manera significativa la vida microbiana (EPA, 1976).

10.1.10. Turbidez.

El último de los parámetros relacionados con la claridad del agua presentó resultados coherentes que ratifican la tendencia descrita en los análisis de la clorofila-a y los sólidos suspendidos totales,

en cuanto a que la obstrucción subóptica del lago se debe a fenómenos biológicos relacionados con el crecimiento y proliferación de algas. De 9.126 pruebas realizadas, 3.277 superaron el umbral, resultando en un 35,9 % de la muestra; es decir, tanto la frecuencia como el alcance afectaron directamente el índice de manera moderada. Sin embargo, el valor máximo de 7,125 sugiere que las excursiones no incidieron en magnitud, sino en cantidad a la amplitud. La tabla (10.10) muestra los resultados del análisis para la turbidez.

TABLA 10.10: Resumen de las pruebas fallidas de la turbidez.

Medida	Valor
Unidades	NTU
Objetivo	5,0
Tipo	Máximo
Mínimo	2,377
Máximo	7,125
Muestra	9.126
Obs. Fallidas	3.277
Porcentaje	35,9

Fuente: Elaboración propia.

El valor objetivo de la turbidez fue adoptado de las especificaciones de la Organización Mundial de la Salud, las cuales están enfocadas en la determinación de estándares en el agua potable. La OMS indica que una medida recomendable debe estar por debajo de 1 NTU; no obstante, este umbral está enfocado en agua potabilizada o tratada de antemano; para cuerpos de agua natural, la medida se flexibiliza hasta un máximo de 5 NTU, indicando que valores altos en la medición de la turbidez pueden proteger bacterias, virus y parásitos dañinos para el ser humano (WHO, 2017).

10.2. Fase 2: Agrupación de los Factores y Cálculo del Índice.

Una vez que se completó el cálculo de la frecuencia y el alcance para cada parámetro, el siguiente paso fue evaluar la amplitud de las pruebas fallidas. La tabla (10.11) resume las excursiones computadas para cada variable. Como ya se había evidenciado en la fase anterior, la clorofila-a obtuvo las mayores excursiones, sumando un total de 6.126,5, lo que representa la lejanía de las estimaciones con respecto al valor objetivo. Nótese como los demás indicadores de claridad en el agua presentaron excursiones moderadas, como es el caso de los sólidos suspendidos totales, con un total de 9,592 y la turbidez, con 379,281. Los parámetros restantes evidenciaron excursiones bajas, debido a que muy pocas pruebas fallaron y a que los valores asociados a dichas pruebas no estuvieron significativamente alejados del umbral guía.

TABLA 10.11: Resumen de las excursiones para el factor de la amplitud.

Parámetro	Total	Fallos	Excursiones	NSE
Arsénico	10.530	0,0	0,0	0,0
Sulfato	7.722	0,0	0,0	0,0
Clorofila-a	13.338	10.043	6.126,49	0,46
Cloruro	8.424	25	0,706	0,000084

Parámetro	Total	Fallos	Excursiones	NSE
Dureza	8.424	66	1,057	0,0000125
pH	9.126	0,0	0,0	0,0
Sodio	12.636	0,0	0,0	0,0
Temperatura	9.828	0,0	0,0	0,0
TSS	12.626	84	9,592	0,000759
Turbidez	9.126	3277	379,281	0,041559

Fuente: Elaboración propia.

Por último, se calculó el índice de calidad del agua para las 702 ubicaciones. El resultado final indicó que 664 puntos tienen una calidad buena y 38, aceptable, como se muestra en la tabla (10.12). Lo que significa que aproximadamente el 95 % de la muestra no se vio afectada en gran medida por los resultados de la clorofila-a; por consiguiente, los demás parámetros contribuyeron principalmente a que la calidad del agua en el lago George se mantuviera estable. En otras palabras, si bien los parámetros individuales de la clorofila-a y la turbidez mostraron excedencias frecuentes de los objetivos ambientales, el índice general clasifica los puntos de análisis como de buena calidad; por lo tanto, la degradación del cuerpo hídrico está principalmente dominada por procesos biológicos sectorizados en tiempo y espacio, más allá de una alteración generalizada en la composición química del líquido.

TABLA 10.12: Resumen de las excursiones para el factor de la amplitud.

Clasificación	Total	Promedio	Máximo	Mínimo	Des. Estándar
Buena	664	84,92	86,97	80,00	1,901
Aceptable	38	75,64	79,98	69,32	2,398

Fuente: Elaboración propia.

Para el correcto análisis de los resultados anteriores, se debe tener en cuenta, de igual forma, la estructura general del índice y la parametrización final de este; es decir, el cálculo considera las medidas que superaron los umbrales por ubicación; por lo tanto, los valores excesivos de la clorofila-a pudieron haberse obtenido en zonas específicas del lago para una época determinada, lo que indica que la claridad del agua puede no verse afectada de manera significativa en un ámbito global, tanto temporal como geográfico; en consecuencia, episodios críticos particulares quedan diluidos por un fenómeno general ideal. Adicionalmente, el hecho de que ocho de los diez parámetros finales utilizados no presentaran fallos, o en su defecto, fallos menores, contribuyó a que el índice se estabilizara.

10.2.1. Distribución del WQI.

Los valores del WQI obtenidos están agrupados principalmente en el rango de 85 a 87 puntos, como lo muestra la figura (10.1), lo que confirma la predominancia de una buena calidad del agua; esto significa que hay excedencias que están afectando directamente la integridad física y química del lago, pero en menor medida, ya sea por ser poco frecuentes o por estar cerca de los valores objetivo. Aún así, los valores más bajos, en el rango de 70 a 80 puntos, indican que hay incumplimientos localizados del índice; sin embargo, la figura (10.2) muestra que las ubicaciones con calidad aceptable no están concentradas en una misma zona, sino que se distribuyen sin un patrón espacial definido sobre el área efectiva de las ubicaciones seleccionadas.

En ese orden de ideas, se puede deducir que la afectación en la calidad del agua no se debe enfocar en el componente geográfico, sino en el temporal; en consecuencia, el fallo recurrente de las pruebas

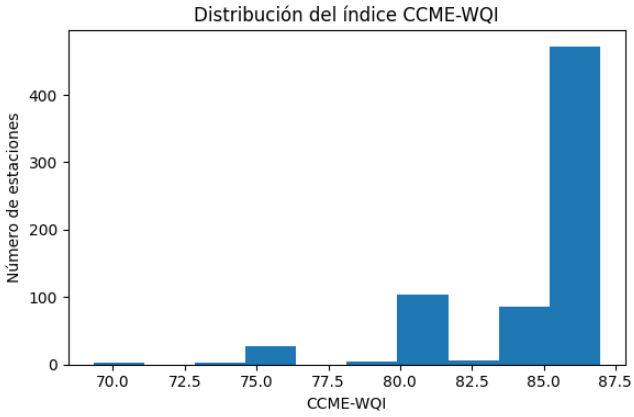


FIGURA 10.1: Distribución del WQI en los puntos evaluables.
Fuente: Elaboración propia.

en los parámetros de claridad está relacionado con el cambio en la composición física del sistema, debido a condiciones estacionales o eventos temporales que afectan la dinámica del lago. Es imperativo aclarar que el análisis anterior involucra solamente las variables que pudieron ser modeladas por medio de una regresión lineal; por ende, es posible que otros parámetros imprescindibles en el análisis de calidad del agua estén repercutiendo en la integridad físico-química del cuerpo y no se estén considerando por la razón anteriormente descrita, como el oxígeno disuelto, la profundidad del disco Secchi (parámetro de claridad), el potasio, el fósforo o el nitrógeno.

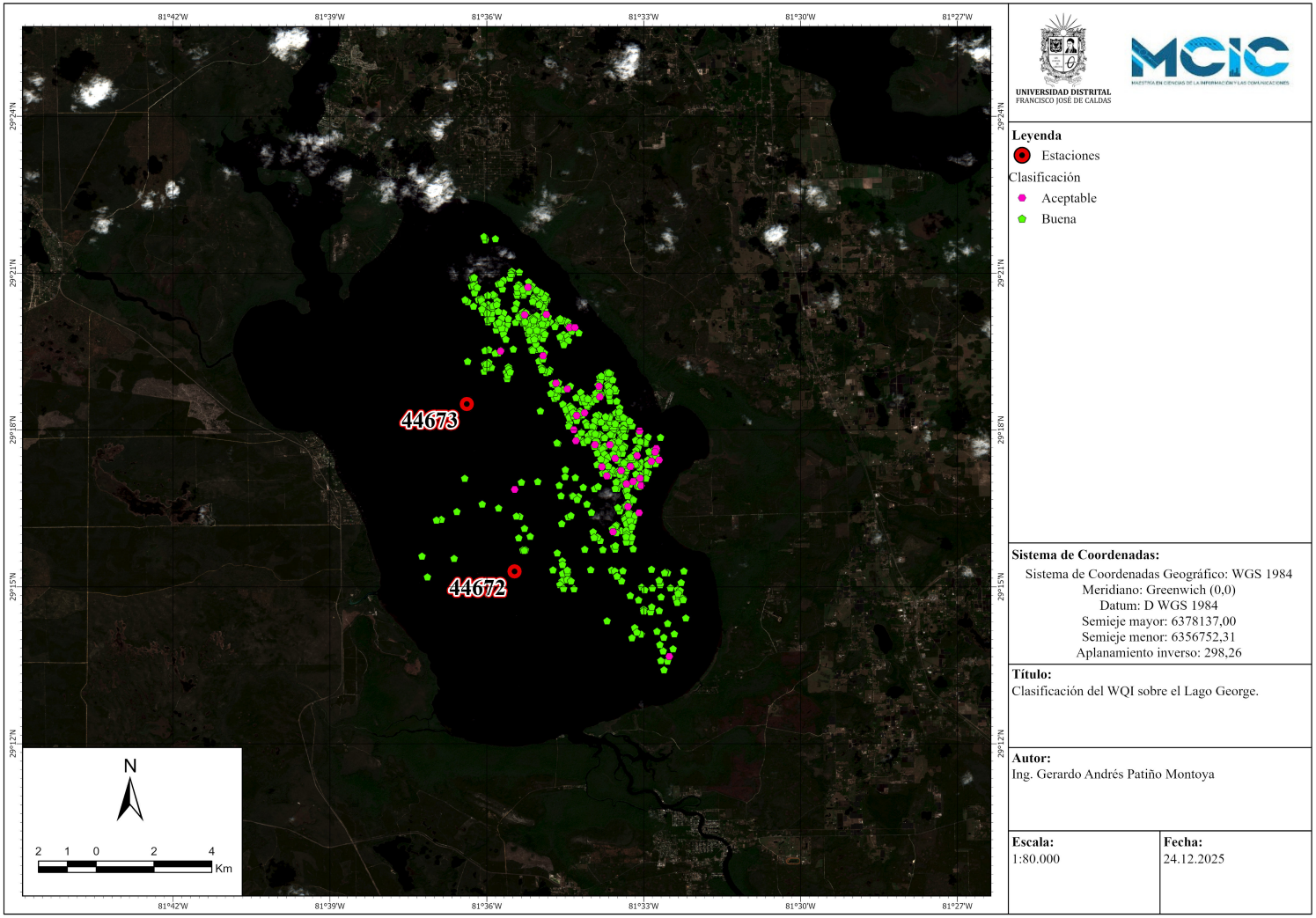


FIGURA 10.2: Clasificación del WQI sobre el lago George.
Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023).

10.3. WQI en las Estaciones de Referencia.

El WQI fue calculado para las dos estaciones de donde se tomaron los parámetros para el año 2023. El resumen de los factores se muestra en la tabla (10.13). El análisis con las medidas reales arrojó resultados ligeramente diferentes en comparación con el índice obtenido a partir de los modelos de regresión lineal. En primer lugar, los parámetros de arsénico, sulfato y sodio no presentaron excedencias, por lo que la interacción de sus factores fue nula, al igual que las estimaciones. El cloruro y la dureza tuvieron varias mediciones que superaron los umbrales en una mayor proporción que las predicciones, con un 5,71 % y 29,27 %, respectivamente; es decir, los minerales tuvieron un protagonismo más importante al castigar significativamente el índice.

Un aspecto diferenciador en el índice de las estaciones estuvo directamente relacionado con la evaluación de la acidez y la temperatura, debido a que las predicciones no mostraron excedencias; no obstante, las mediciones en campo sí lo hicieron. Por parte del pH, el 14,52 % de la muestra falló; para la temperatura, el 2,55 %. Si bien los porcentajes son bajos, el hecho de que los umbrales se hayan superado ya representa una activación directa de los tres factores, siendo la frecuencia la que más peso tuvo al inferir que todo el parámetro falló con solo un registro no válido. En ese orden de ideas, las estaciones fallaron en más parámetros que los modelos, lo que sugiere que el WQI será menor.

Finalmente, los parámetros relativos a la claridad del agua fueron bastante similares. Los sólidos suspendidos totales de las estaciones no fallaron; sin embargo, las estimaciones sí lo hicieron en un 0,7 % de la muestra total. La turbidez falló sustancialmente en aproximadamente un tercio de la muestra para ambos análisis. Por último, la clorofila-a presentó excedencias en el 75,3 % de las predicciones y en el 76,12 % de las mediciones. En síntesis, los parámetros que más incidencia tuvieron en el WQI presentaron dinámicas similares tanto en las regresiones como en las estaciones. Lo que confirma la confiabilidad y robustez de los modelos finales elegidos y las subsecuentes predicciones.

TABLA 10.13: Resumen de las excursiones del WQI en las estaciones de referencia.

Parámetro	Total	Fallos	Porcentaje	Excursiones	NSE	Mínimo	Máximo	Umbral
Arsénico	35	0,0	0,0 %	0,0	0,0	0,52	0,92	1,0
Sulfato	35	0,0	0,0 %	0,0	0,0	16,03	66,94	250,0
Cfila-a	67	51	76,12 %	42,229	0,630	3,73	48,46	15,0
Cloruro	35	2	5,71 %	0,118	0,0034	77,34	264,79	250,0
Dureza	41	12	29,27 %	1,007	0,026	69,98	208,86	180,0
pH	241	35	14,52 %	0,355	0,0015	7,16	8,68	6,5 - 8,5
Sodio	33	0,0	0,0 %	0,0	0,0	40,51	166,71	200,0
Temp.	157	4	2,55 %	0,084	0,0005	56,12	87,134	86,0
TSS	39	0,0	0,0 %	0,0	0,0	3,6	20,60	25,0
Turbidez	38	12	31,58 %	0,283	0,074	1,43	7,33	5,0

Fuente: Elaboración propia.

La integración de los factores en el índice final para las estaciones indicó que la calidad del agua en la estación 44762 es buena, dado un valor de 81,78; mientras que en la estación 44763 la calidad es mala, al presentar un WQI de 63,75, como se evidencia en la tabla (10.14). Este resultado permite concluir que las mediciones que fallaron en los parámetros adicionales, en comparación con los modelos, afectaron en mayor medida a esta última estación, hasta el punto de reducir el valor asociado a la penúltima categoría; aún así, dicho valor se ubicó bastante cerca del umbral para considerarse aceptable, exactamente a 1,25 puntos. Asimismo, la distancia al valor mínimo obtenido con las estimaciones

fue de 5,57 unidades. En síntesis, si bien la estación 44673 difiere notablemente en la categorización, los valores continuos indicaron que el comportamiento en ambas estaciones no está tan distante del escenario estimado por los modelos, condicionado igualmente por la cantidad de observaciones con las que se evaluó el índice final.

TABLA 10.14: WQI calculado para las estaciones.

Clasificación	**Estación**	**WQI CCME**
Buena	44672	81,78
Mala	44673	63,75

Fuente: Elaboración propia.

En definitiva, se obtuvieron 702 ubicaciones en las cuales los modelos fueron evaluados, más dos estaciones de referencia, para un total de 704 puntos de cálculo efectivo. Allí prevaleció que la calidad del agua en el lago George es principalmente buena, afectada en mayor medida por elementos eutróficos relativos al desarrollo biológico de algas y vegetación acuática en general, siendo la clorofila-a y la turbidez los parámetros con mayor protagonismo en el índice. Adicionalmente, se presentaron zonas con calidad aceptable sin un patrón espacial determinado, por lo que la disminución del valor final se atribuyó a dinámicas temporales particulares y no a un fenómeno geográfico evidente.

11

Implementación del Modelo MARS.

El procedimiento desarrollado en los capítulos anteriores determinó un valor numérico continuo para categorizar la calidad del agua, según la metodología del Consejo Canadiense de Ministros de Medio Ambiente, y con base en el uso de predicciones estimadas con modelos de regresión lineal en puntos de evaluación válidos sobre la superficie del lago George, en un periodo comprendido entre septiembre de 2022 y abril de 2024. En el presente apartado, se desarrolló posteriormente la implementación de la regresión de splines adaptativos multivariada con el fin de modelar el índice a partir de valores de reflectancia determinados por cada banda del programa Sentinel-2.

11.1. División de Épocas Estacionales.

La base de datos final con la que se ajustó el modelo MARS reunió un total de 704 observaciones con un solo índice por registro; es decir, ya no se tuvo en cuenta un componente temporal explícito, sino solo geográfico. Esta aclaración tiene varias implicaciones en la medida en que el presente análisis no pretendió evaluar qué imagen explica mejor el WQI, sino que buscó interpretar qué estado espectral del lago fue más representativo en el índice integrado para el periodo en cuestión. En consecuencia, se eligieron cuatro imágenes que representarían cada estado estacionario (invierno, primavera, verano y otoño), de las cuales se extrajeron los valores de reflectancia de las 12 bandas para cada uno de los puntos de análisis de forma independiente.

La elección de las imágenes se realizó hallando cuáles de estas no habían sido removidas en el proceso de obtención de las ubicaciones evaluables para todos los parámetros, siendo febrero, marzo, septiembre, octubre y noviembre de 2023, y abril de 2024 las únicas que presentaron puntos válidos para todos los modelos. Adicionalmente, octubre y diciembre de 2022, diciembre de 2023 y enero y febrero de 2024 solo fueron eliminadas una vez durante el proceso de filtrado. Por consiguiente, para representar el comportamiento del lago en invierno se eligió la imagen correspondiente al mes de diciembre de 2022; para la época de primavera, se seleccionó la imagen de marzo de 2023; el verano fue explicado por la imagen de julio; finalmente, para el otoño se optó por la imagen de octubre del mismo año.

11.1.1. Invierno.

De antemano, cada muestra fue dividida en un 70 % para la fase de entrenamiento, dejando el restante 30 % de la muestra para la fase de prueba. Asimismo, para analizar el desempeño de los modelos, se procedió a calcular la raíz del error medio cuadrático (RMSE), el error medio absoluto (MAE), el error relativo y el coeficiente de determinación R^2. La regresión para el mes de diciembre presentó un comportamiento que se iba a repetir de manera constante a lo largo del ejercicio. En primer lugar, se obtuvo un RMSE de 2,83, un MAE de 2,01 y un error relativo de 2,45, como se evidencia

en la tabla (11.1); esto indicó que los residuos mostraron una variabilidad relativamente baja y que las medidas de desempeño estaban bastante cerca entre sí con valores bajos.

TABLA 11.1: Desempeño del modelo MARS en diciembre.

Medida	**Valor**
RMSE	2,83
MAE	2,01
Error relativo	2,45
R^2	-0,042

Fuente: Elaboración propia.

No obstante, el problema real surgió con la interpretación del coeficiente de determinación, debido a que arrojó un valor de -0,042. Lo cual indicó que la variabilidad del WQI no pudo ser explicada por las bandas. A priori, los resultados parecerían entrar en una contradicción, puesto que el modelo tuvo un buen rendimiento desde el punto de vista predictivo, pero no logró explicar de manera significativa la dinámica específica de la variable independiente. En otras palabras, dada una variabilidad tan baja del índice en el lago, el modelo fue capaz de estimar un valor medio sobre una muestra estrecha con baja dispersión; sin embargo, no logró capturar una estructura espacial o espectral compleja.

Para explicar por qué el R^2 es tan bajo, obsérvese la estructura general del coeficiente dada por la ecuación (11.1), en la cual se obtiene una relación directa entre el R^2 y la variabilidad del modelo; es decir, cuanto más grande sea el grado de dispersión de la variable respuesta con respecto a los residuos, más grande será el coeficiente de determinación; por lo tanto, si la variable explicada presenta una variabilidad baja, el R^2 tenderá a bajar de igual forma, por más que los residuos también sean bajos. En otras palabras, el coeficiente colapsa debido a la varianza tan pequeña del WQI.

$$R^2 = 1 - \frac{\sigma_r^2}{\sigma_m^2} \tag{11.1}$$

Donde σ_r^2 es la varianza de los residuos y σ_m^2 es la varianza del modelo; en este caso, del índice.

11.1.2. Primavera.

El modelo ajustado con los valores de reflectancia de la imagen correspondiente al mes de marzo resultó estar alineado con el desempeño del mes de diciembre, sin cambios abruptamente drásticos. El error medio cuadrático varió en solo 0,02 puntos, siendo ligeramente más alto en primavera. Un comportamiento similar se evidenció en el error medio absoluto y el error relativo, estando estos dos parámetros muy cerca de los resultados de invierno, pero por encima en todo momento. El coeficiente de determinación, una vez más, resultó bastante bajo, incluso negativo. Ante este panorama, se pudo concluir que el índice no estuvo fuertemente influenciado por temperaturas bajas y por condiciones asociadas a las dos primeras estaciones del año. El resumen de las estadísticas para el modelo MARS en el mes de marzo se presenta en la tabla (11.2).

TABLA 11.2: Desempeño del modelo MARS en marzo.

Medida	Valor
RMSE	2,85
MAE	2,05
Error relativo	2,50
R^2	-0,053

Fuente: Elaboración propia.

11.1.3. Verano.

Teniendo en cuenta que los parámetros que tuvieron una mayor incidencia en el índice fueron la clorofila-a y la turbidez, era de esperarse que la época en la que más actividad eutrófica se presentara fuera aquella que mostrara los resultados con menor dispersión; en otras palabras, un mejor desempeño. La hipótesis preliminar se pudo confirmar con los resultados de la tabla (11.3), en el cual se puede evidenciar que todas las medidas para el mes de junio estuvieron por debajo de los valores obtenidos en los escenarios anteriores; no obstante, en términos prácticos, la distancia entre estas no fue notable; por ejemplo, el RMSE solo bajo 0,06 puntos, al igual que el MAE. El R^2, por otra parte, si presentó un cambio más evidente al pasar de -0,053 a -0,013; aún así, el ajuste en este parámetro sigue siendo pobre, lo que indica que el modelo sigue prediciendo un valor central y no está contemplando completamente una estructura especial o espectral más compleja.

TABLA 11.3: Desempeño del modelo MARS en julio.

Medida	Valor
RMSE	2,79
MAE	1,99
Error relativo	2,43
R^2	-0,013

Fuente: Elaboración propia.

11.1.4. Otoño.

El mes de octubre presentó el peor ajuste dentro de la escasa variabilidad y cercanía de las medidas entre sí. El error medio cuadrático fue de 2,85, similar al mes de marzo; el error medio absoluto fue el mayor de todas las épocas evaluadas, al igual que el error relativo, con valores de 2,11 y 2,56, respectivamente. Con los resultados de la tabla (11.4), y los descritos anteriormente, se evidenció que la actividad espectral a lo largo del periodo de estudio no presenta un grado de dispersión demasiado grande, estando las medidas de desempeño considerablemente bajas y cercanas para los cuatro meses en análisis.

TABLA 11.4: Desempeño del modelo MARS en octubre.

Medida	Valor
RMSE	2,85
MAE	2,11
Error relativo	2,57

Medida	Valor
R^2	-0,054

Fuente: Elaboración propia.

En síntesis, gracias a la homogeneidad obtenida con los modelos de regresión lineal, el modelo MARS no logró estructurar un sistema complejo que relacionara los componentes geográfico y espectral, lo cual no significó que el modelo fuera ineficiente o inutilizable; sino que, debido a la naturaleza propia del índice en la zona de estudio, se hizo muy difícil implementar un modelo que pudiera detectar gradientes abruptos en la calidad del agua o explicar variaciones sectorizadas del índice; por otra parte, el modelo sí podría ser utilizado para confirmar una homogeneidad espacial en el diagnóstico inicial de la calidad e iniciar estudios exploratorios en ubicaciones donde las estimaciones estén fuera de rango o presenten resultados discrepantes.

11.2. Mejor Modelo.

11.2.1. Primera iteración.

De acuerdo con el resumen presentado para cada mes, se demostró que la incidencia espectral en el índice no fue tan variable a lo largo del año; aun así, se eligió la tercera regresión, es decir, el mes de julio, como el mejor modelo. La estructura general se evidencia en la ecuación (11.2), en la cual intervinieron solamente las bandas 3 (verde), 6 (borde de vegetación rojo 2), 9 (vapor de agua) y 11 (infrarrojo de onda corta 1). Las demás bandas fueron descartadas durante el proceso de filtrado del MARS.

$$\begin{aligned} WQI = 84,56 &- 1.195 * \max(0; B_3 - 0,024) \\ &- 18.802 * B_9 * \max(0; 0,0024 - B_3) \\ &+ 12.507 * (B_{11} * B_6) \end{aligned} \tag{11.2}$$

La banda del verde es la que más incidencia tuvo, puesto que estuvo presente en dos funciones bisagra de maneras distintas: en el primer término, luego del intercepto, el índice disminuye fuertemente cuando el nivel valor de reflectancia de esta banda es mayor a 0,024, lo que se traduce en una peor calidad del agua cuando los elementos asociados a la claridad del agua, como la clorofila-a, la turbidez y los TSS, interactúan activamente con la banda 3; por otra parte, cuando los niveles valores de reflectancia del verde son menores que 0,024, la función bisagra interactúa con la banda 9, la cual penaliza el índice cuando los elementos asociados al vapor de agua interactúan activamente con el sensor; en otras palabras, cuando la banda 3 presenta valores pequeños, se activa la banda 9, la cual penaliza el WQI mientras más alto sea su valor espectral.

Por último, el modelo identificó una función adicional que relaciona las bandas 11 y 6, cuya interacción es directa con la calidad del agua; es decir, el aumento de los valores de reflectancia en cualquiera de las dos bandas hace que el índice aumente, por lo que se puede interpretar que niveles altos en el infrarrojo cercano de onda corta y el borde de vegetación rojo están asociados con una buena calidad del agua. Cabe resaltar que la longitud de onda de estas bandas está bastante alejada del espectro visible; por tanto, el alcance y sensibilidad se relacionan con fenómenos y dinámicas que se desarrollan en la profundidad del lago, no tanto en la superficie.

11.2.2. Segunda iteración.

El modelo MARS involucra una gran cantidad de funciones bisagra al hacer el ajuste; dichas funciones relacionan no solo las bandas finales seleccionadas, sino también combinaciones de las bandas restantes, generando ruido que afecta la suma de cuadrados de los residuos. Por tal motivo, se decidió ajustar nuevamente el modelo, pero solamente con las bandas implicadas en la ecuación (11.2), con el fin de reducir las posibles perturbaciones generadas por las demás variables independientes. Las medidas de desempeño que se presentan en la tabla (11.5) mostraron una leve mejoría en el error medio cuadrático y en el coeficiente de determinación, lo que confirmó que las bandas restantes sí estaban induciendo ruido en los residuos, provocando una mayor variabilidad en la suma de los cuadrados; sin embargo, esto no afectó el desempeño de las estimaciones y sus medidas.

TABLA 11.5: Desempeño del modelo MARS en julio solo con las bandas más significativas.

Medida	**Valor**
RMSE	2,76
MAE	1,99
Error relativo	2,43
R^2	0,009

Fuente: Elaboración propia.

Adicionalmente a la reducción del RMSE y al leve aumento en el R^2, la ecuación generada por la segunda iteración desestimó las bandas 6 y 11 en el ajuste final, como se aprecia en la ecuación (11.3). Además, el modelo identificó otro punto de corte, esta vez para la banda 9, en el valor de 0,0138, el cual se encuentra a 0,0043 unidades del promedio de la banda en mención. Una tercera iteración fue ajustada solamente con las bandas 3 y 9; no obstante, los resultados no mejoraron con respecto a la segunda; por esta razón, y con base en el principio de parsimonia, se decidió proceder con este modelo para su posterior evaluación y validación.

$$\begin{aligned} WQI = {} & 85,2848 - 152,601 * \max(0; 0,024 - B_3) \\ & + 38435,6 * \max(0; B_9 - 0,0138) * \max(0; B_3 - 0,024) \\ & - 2486,56 * \max(0; B_9 - 0,0138) \end{aligned} \tag{11.3}$$

11.2.3. Dominio espectral.

Un proceso clave en la implementación de cualquier modelo de regresión es la determinación del dominio; es decir, el rango de valores de las variables independientes en el cual el modelo es evaluable y permite generar estimaciones válidas. Para el modelo final de julio, se obtuvo una serie de estadísticas descriptivas con el fin de interpretar las condiciones con las que se ajustaron las funciones bisagra y su relación con las predicciones, tal como se hizo con los modelos de regresión lineal de los parámetros.

La banda 3 presentó el valor más alto entre todas las que participaron en el modelo final. El promedio fue de 0,018, el cual está a una y media desviación estándar de distancia con respecto al umbral de las dos primeras funciones bisagra. El dato anterior es importante en la medida en que la regresión no inventó puntos de corte ficticios, por lo que el modelo es evaluable de una manera factible. Los rangos indican una variabilidad suficiente al abarcar un espectro considerablemente amplio de predicción. El resumen completo de las estadísticas descriptivas se presenta en la tabla (11.6).

TABLA 11.6: Estadísticas descriptivas de las bandas del MARS para el mes de julio.

Banda	**Mínimo**	**Máximo**	**Rango**	**Media**	**Desv. Estándar**
3	0,0110	0,0870	0,0760	0,0180	0,0041
6	0,0086	0,0825	0,0739	0,0127	0,0031
9	0,0034	0,0678	0,0644	0,0095	0,0028
11	0,0044	0,0703	0,0659	0,0058	0,0027

Fuente: Elaboración propia.

Si bien las bandas 6 y 11 no terminaron siendo consideradas en la ecuación final, siguen estando relacionadas en la regresión debido a que el modelo MARS involucra todas las variables en el ajuste y las posteriores estimaciones; por tal motivo, se mencionaron los rangos y las medidas descriptivas en la tabla anterior.

12

Evaluación del Modelo MARS.

La evaluación del modelo MARS funciona de manera similar al uso de un modelo de regresión lineal ordinario; es decir, se evalúan los valores de las variables independientes en la ecuación final, con la salvedad de que esta metodología involucra una serie de validaciones proporcionadas por las funciones bisagra y los puntos de corte, los cuales activan o desactivan los estimadores de acuerdo con dichos valores. Esta etapa del proceso consistió en una única fase repetida para tres escenarios diferentes, en la cual se validó que los valores de las bandas involucradas cumplieran con los rangos establecidos en la ecuación (11.3).

En primer lugar, se utilizó la muestra de 100.000 puntos aleatorios generada para la determinación de los puntos evaluables en los modelos de los parámetros de calidad del agua; luego, se creó una máscara efectiva en la zona con menor densidad de ubicaciones, la cual fue la zona occidental del lago. Posteriormente, se extrajeron los puntos que estuvieran dentro del área efectiva de evaluación y se validó que los valores de reflectancia de dichas ubicaciones estuvieran dentro de los rangos mencionados anteriormente.

No obstante, se presentó un problema crítico que contribuyó a la explicación de por qué el cálculo del WQI solo se pudo realizar en la zona oriental del lago. Obsérvese en la figura (12.1) cómo la presencia de nubes y perturbaciones generales en el agua altera la superficie observable del lago, haciendo que el sensor capture valores espectrales más altos, es decir, más claros; esto influye negativamente en la validación de los puntos evaluables; aún así, se logró obtener una muestra considerablemente robusta para analizar el desempeño del modelo en una zona con características ligeramente diferentes sobre la misma superficie del lago.

12.1. Escenario 1: Julio.

La muestra total de puntos válidos para el mes de julio fue de 40.512 puntos evaluables, generando los resultados que se presentan en la tabla (12.1). Para las distintas combinaciones de los valores de las bandas 3 y 9, no se obtuvieron índices que identificaran zonas con calidad excelente; sin embargo, el 10 % de las estimaciones tuvo un valor de 44,9 o menor, indicando una calidad muy mala en estas áreas. Adicionalmente, el número de predicciones con categoría mala superó al de aquellas con calidad aceptable, un fenómeno que no se había observado anteriormente con los datos de entrenamiento.

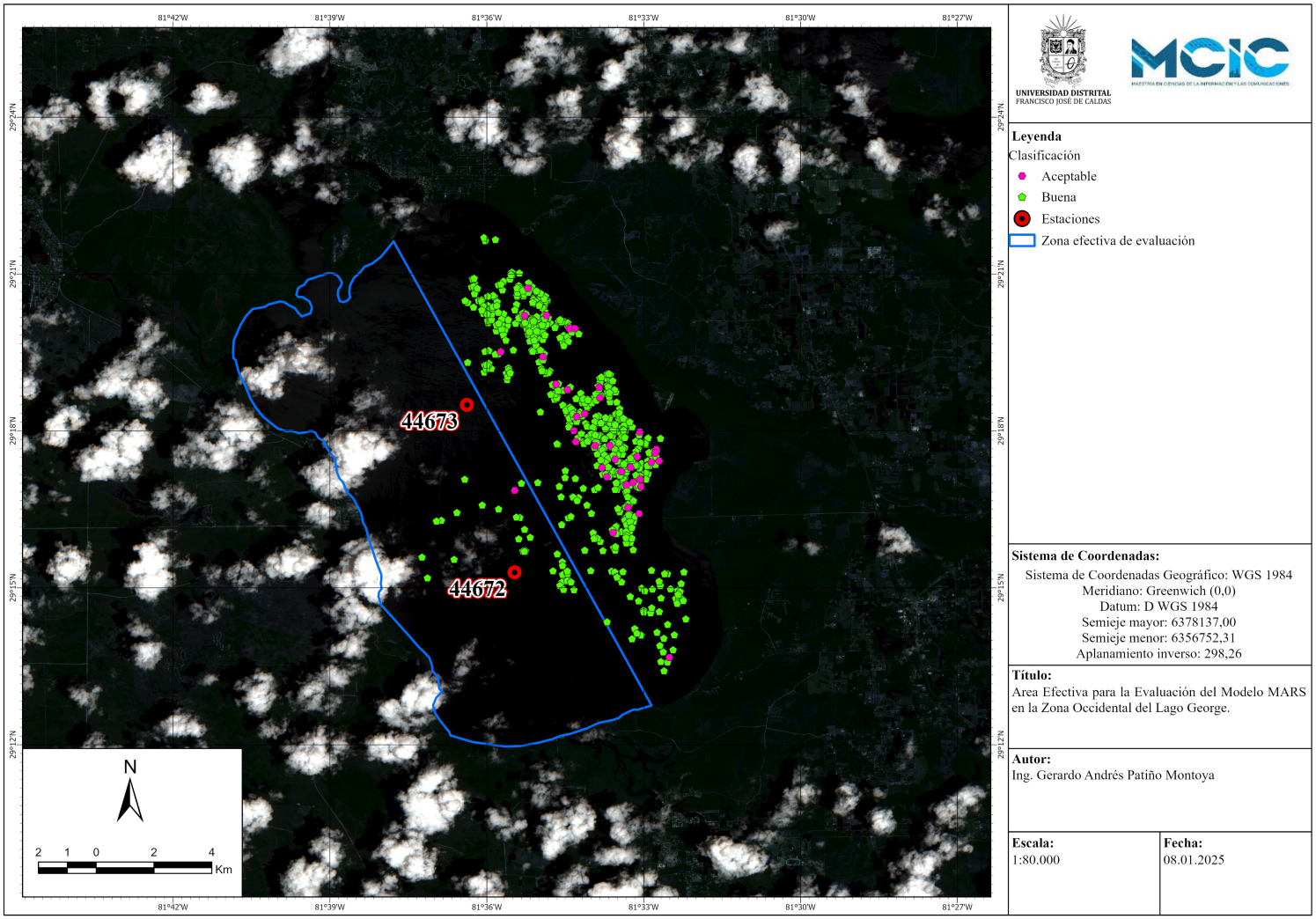


FIGURA 12.1: Área efectiva para la evaluación del modelo MARS en la zona occidental del lago George.
Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023).

TABLA 12.1: Resumen de la evaluación del modelo MARS en la zona occidental del lago George para el mes de julio.

Categoría	Cantidad	Porcentaje
Buena	24.048	48 %
Aceptable	6.707	19 %
Mala	8.139	23 %
Muy mala	3.748	10 %
Fuera de rango	281	Menos del 1 %

Fuente: Elaboración propia.

Los resultados de la evaluación indican una clara extrapolación local del MARS en escenarios con actividad espectral diferente de la representada en la etapa de entrenamiento, condicionada también por la alta presencia de nubes y perturbaciones generales sobre la superficie lacustre. Esto no significa que el modelo sea inválido; por el contrario, permite analizar las diferentes dinámicas del lago a lo largo de su extensión. La heterogeneidad en las estimaciones también se debe a una identificación de gradientes de una manera más continua, dado que los saltos de evaluación son mucho menores.

Igualmente, un aspecto crítico de esta etapa del procesamiento estuvo relacionado con la obtención de valores de calidad mala y muy mala (prácticamente un tercio de la muestra). Esto apuntó a que el modelo en la fase de predicción no estaba implementando una metodología basada en promedios, sino que estaba extrapolando las estimaciones para capturar fenómenos de contrastes espectrales más fuertes; es decir, los umbrales locales o puntos de corte alteran drásticamente la estructura de las funciones bisagra ante valores de reflectancia cercanos a los límites de los rangos (ya sean superiores o inferiores) de cada variable independiente, lo que provocó que el modelo generara índices mucho menores que los esperados.

Por otra parte, la distribución espacial de los índices refleja un comportamiento específico más revelador. Obsérvese en la figura (12.2) cómo la presencia de nubes cubre la superficie occidental del lago; se puede evidenciar claramente que una gran cantidad de observaciones de calidad muy mala, mala y aceptable se encuentra condensada alrededor de las zonas con valores de reflectancia más altos, lo que sugiere que el desempeño del modelo se vio afectado por los contrastes generados entre la superficie nubosa y el lago.

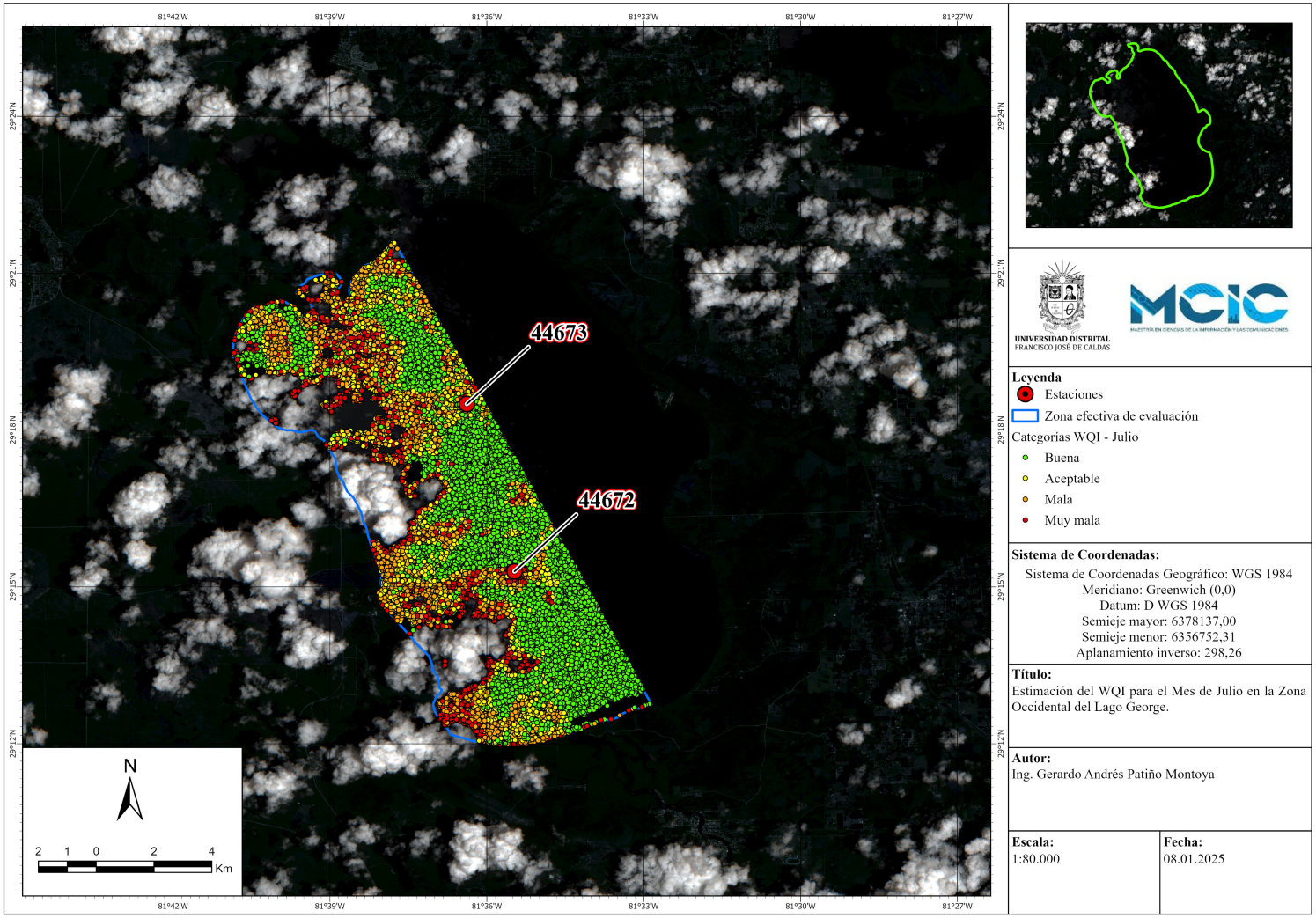


FIGURA 12.2: Estimación del WQI para el mes de julio en la zona occidental del lago George.
Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023).

12.2. Escenario 2: Agosto.

Teniendo en cuenta los resultados del primer escenario, se decidió evaluar el modelo MARS en una imagen con características estacionales y espectrales similares; por tal motivo, se tomó la imagen capturada en el mes de agosto y se repitió el proceso explicado anteriormente. La validación de los puntos arrojó un total de 12.690 ubicaciones válidas, una reducción significativa en comparación con el mes anterior; no obstante, sigue siendo representativa para efectos del ejercicio.

La dinámica del desempeño del modelo en el segundo escenario reflejó un comportamiento muy parecido respecto al del primero, en la medida en que un 43 % de la muestra se categorizó como de buena calidad, es decir, una reducción porcentual del 5 %, como se evidencia en la tabla (12.2). Las zonas con calidad aceptable aumentaron ligeramente en términos porcentuales; no obstante, las categorías de calidad mala y muy mala siguieron predominando en más de un tercio de la muestra. Esto demuestra que el modelo siguió extrapolando en zonas con una actividad espectral más irregular, coaccionada por la presencia de nubes, las cuales generaban picos en los valores de reflectancia de las bandas, provocando una reducción en la calidad general del lago.

TABLA 12.2: Resumen de la evaluación del modelo MARS en la zona occidental del lago George para el mes de agosto.

Categoría	Cantidad	Porcentaje
Buena	5.448	43 %
Aceptable	2.711	21 %
Mala	3.581	28 %
Muy mala	821	7 %
Fuera de rango	129	1 %

Fuente: Elaboración propia.

Visualmente, la dinámica espectral de las nubes siguió siendo un problema, lo que confirmó la hipótesis propuesta en la evaluación del modelo para el mes de julio, la cual indicaba que la presencia de nubes tiene efectos negativos directos en las estimaciones del modelo, en la medida en que los píxeles más cercanos a las superficies más claras; es decir, las zonas de transición entre la parte nubosa y el lago reflejan una gran cantidad de energía, lo que se traduce en valores espectrales mucho más altos, que a su vez se encuentran en el límite de los intervalos de evaluación del modelo.

Esto implicó que el modelo generara predicciones en condiciones para las cuales no fue entrenado, dado que la validación de los puntos de evaluación en los modelos de regresión lineal de los parámetros filtró aquellas zonas con valores de píxel más elevados; por lo tanto, tanto las nubes cómo las perturbaciones generadas por la sombra y el movimiento de estas sobre la superficie del lago provocaron que el modelo subestimara las áreas con mayor actividad espectral, reduciendo las estimaciones, incluso, a resultados negativos o muy cercanos a cero. En consecuencia, fue necesario evaluar el modelo en una imagen cuyas condiciones meteorológicas fueran lo más homogéneas posible, de tal forma que la evaluación del modelo se hiciera en un ambiente más parecido al de la etapa de entrenamiento y prueba.

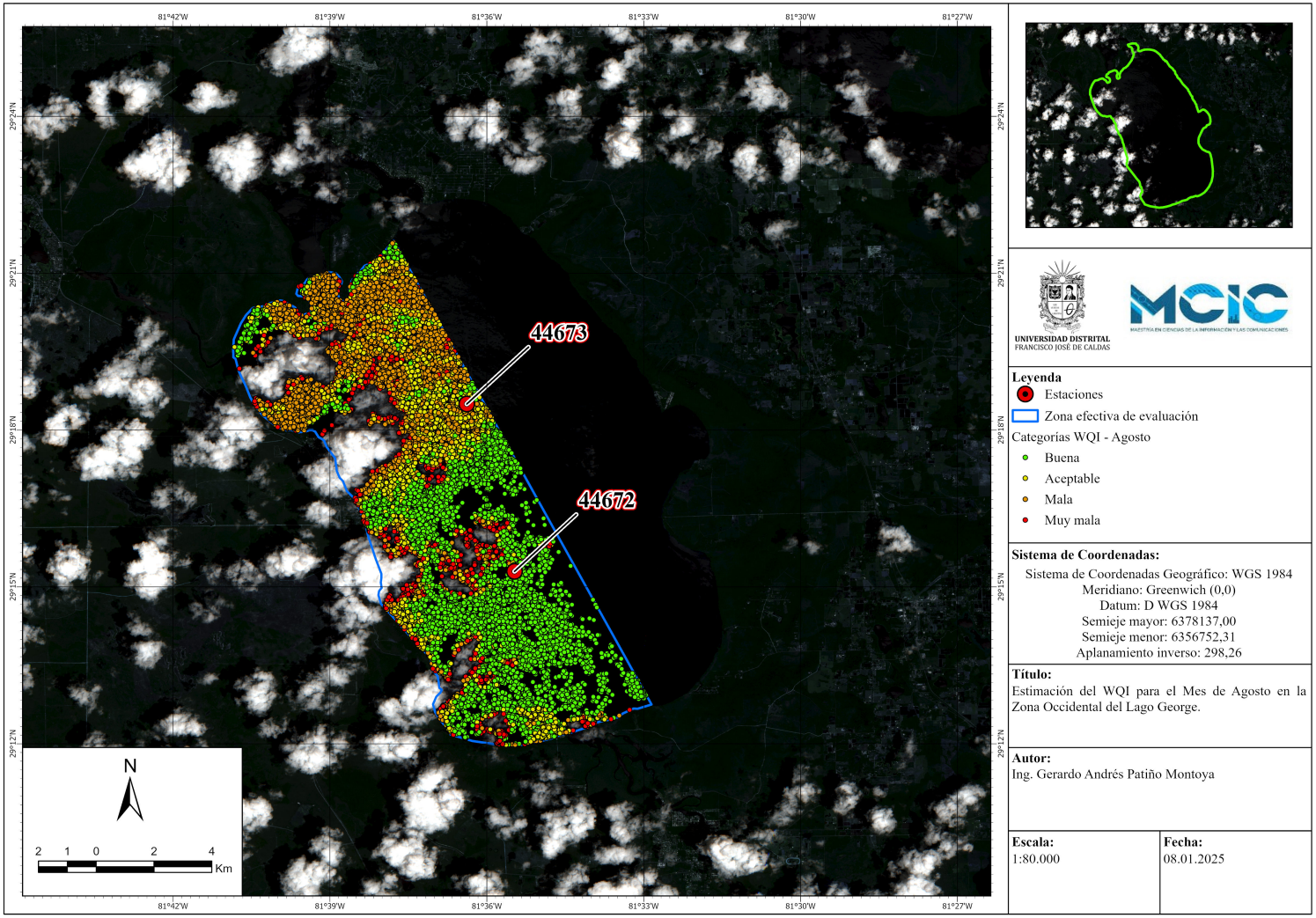


FIGURA 12.3: Estimación del WQI para el mes de agosto en la zona occidental del lago George.
Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023).

12.3. Escenario 3: Septiembre.

Los primeros dos escenarios permitieron deducir que las nubes afectaban notablemente el desempeño del modelo, provocando que una porción considerable del área de evaluación presentara valores mínimos y negativos del índice estimado. Ante este panorama, se seleccionó una imagen adicional que permitiera corroborar los hallazgos descubiertos, pero en una superficie más homogénea, sin las implicaciones de la cobertura en mención. Afortunadamente, las condiciones meteorológicas para el mes subsiguiente mejoraron significativamente, permitiendo la plena identificación de la zona occidental en la parte norte y sur del lago. Lo que no solo significó un análisis estacional idóneo, sino que también se conservó la integridad y la consistencia temporal al tratarse de escenarios continuos.

La validación de las ubicaciones filtró un total de 23.456 puntos de evaluación válidos, cerca del doble en comparación con el escenario del mes de agosto. Las estimaciones sobre la muestra seleccionada arrojaron resultados certeros con relación a las hipótesis elaboradas en los dos primeros escenarios. En primer lugar, la categoría de calidad buena prevaleció firmemente en las predicciones, representando un 96 % del total de observaciones; por consiguiente, el porcentaje restante se distribuyó entre las categorías de aceptable, mala, muy mala y los valores fuera de rango. El resumen de la evaluación del modelo sobre la imagen de septiembre se muestra en la tabla (12.3).

TABLA 12.3: Resumen de la evaluación del modelo MARS en la zona occidental del lago George para el mes de septiembre.

Categoría	**Cantidad**	**Porcentaje**
Buena	22.580	96 %
Aceptable	258	1 %
Mala	331	1 %
Muy mala	163	Menos del 1 %
Fuera de rango	124	Menos del 1 %

Fuente: Elaboración propia.

En cuanto a la distribución espacial de los puntos evaluados, se puede apreciar que hubo una porción principal agrupada en la zona noroccidental del lago, lo que supuso un escenario óptimo, puesto que particularmente esta zona se encontraba contaminada (en términos espectrales) con coberturas no deseadas en los análisis anteriores. Adicionalmente, se identificaron algunos puntos extra en la zona sur, muy cerca del límite de desembocadura del afluente del río San Juan Bautista, el cual representa uno de los sectores de abastecimiento más importantes del lago, debido a que dicho río fluye de sur a norte.

A partir de la figura (12.4), se pudo evidenciar claramente que el modelo ya fue capaz de generar predicciones más estables en las zonas identificadas como críticas, permitiendo validar la homogeneidad de la calidad del agua en el lago en condiciones climáticas estables. Si bien la obtención de valores atípicos extremos, como aquellos fuera de rango o asociados a la categoría más baja del índice, fue constante en el presente escenario, la predominancia de estimaciones más acordes a los valores de entrenamiento simboliza la estabilización y confiabilidad sobre las predicciones, permitiendo confirmar que los parámetros que más afectan la calidad del agua en el cuerpo se deben a una actividad biológica y eutrófica general y no a un fenómeno temporal o geográfico focalizado.

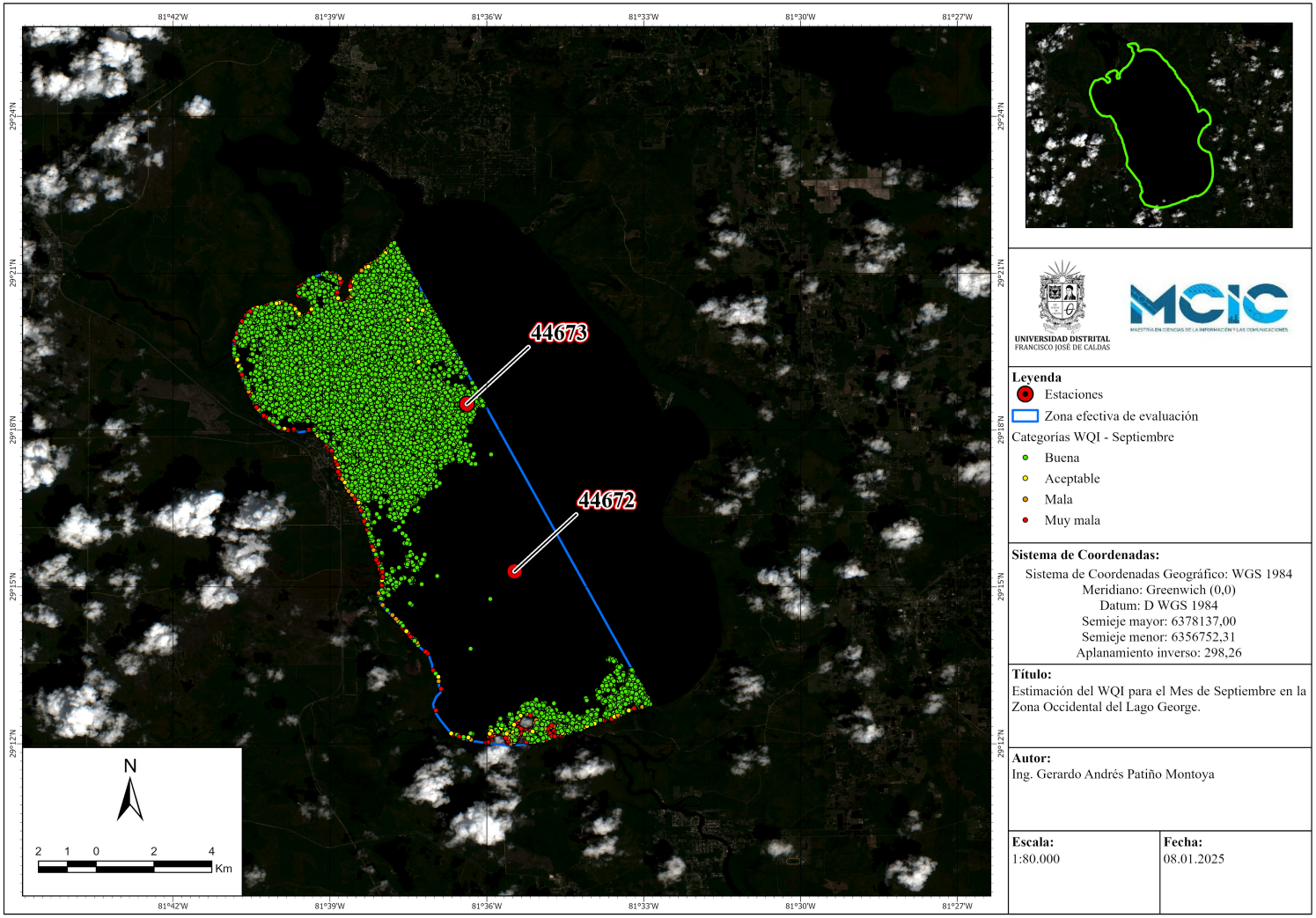


FIGURA 12.4: Estimación del WQI para el mes de septiembre en la zona occidental del lago George.
Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023).

12.4. Resumen de los Escenarios.

La evaluación del modelo permitió identificar varios aspectos importantes del comportamiento espectral, tanto del lago como de la regresión, tales como la limitación de la implementación ante condiciones atmosféricas heterogéneas, asociadas con la presencia de nubes y perturbaciones en la superficie del lago. Esto se debió a que la lectura espectral del sensor capturaba valores de reflectancia elevados, lo que ocasionaba una subestimación extrema de las predicciones; no obstante, se pudo corroborar que el desempeño del modelo se estabiliza cuando se presentan condiciones meteorológicas más estables, provocando que el MARS estime valores del WQI más realistas, acorde con la situación real del agua.

En la tabla (12.4) se resumen las estadísticas descriptivas de las estimaciones por categoría para cada escenario. Inicialmente, se evidenció una constancia en los valores mínimos y máximos para cada clase, lo que significó que el modelo consideró todo el espectro activo en lugar de concentrar los resultados alrededor de un valor promedio por categoría. En segundo lugar, ninguna predicción estuvo por encima de los 86 puntos en la escala del índice; por lo que el modelo solo presentó problemas de subestimación y no de sobrestimación. En tercer lugar, la variabilidad de las predicciones aumenta a medida que el índice se reduce, lo cual era esperable, ya que la fase de entrenamiento y prueba se realizó sobre datos de calidad buena, aceptable y mala (en un solo registro); por lo tanto, aquellas superficies que exigen al modelo con valores extremos en el rango evaluable ocasionan que las estimaciones presenten cifras físicamente inverosímiles.

TABLA 12.4: Estadísticas descriptivas de la evaluación del modelo MARS en los tres escenarios propuestos.

Escenario	Categoría	Mínimo	Máximo	Promedio	Des. Estándar
Julio	Buena	80,00	85,29	84,16	0,96
	Aceptable	65,00	79,99	72,47	4,33
	Mala	45,00	64,99	55,99	5,66
	Muy Mala	0,00	44,99	29,67	13,57
Agosto	Buena	80,00	85,28	84,28	1,02
	Aceptable	65,00	79,99	72,20	4,29
	Mala	45,01	64,99	56,70	5,01
	Muy Mala	0,00	44,99	24,59	15,75
Septiembre	Buena	80,00	85,28	84,34	0,50
	Aceptable	65,06	79,85	73,47	4,61
	Mala	45,10	64,99	56,75	5,85
	Muy Mala	0,00	44,99	16,42	16,19

Fuente: Elaboración propia.

En resumen, el modelo presenta un rendimiento esperable, el cual genera estimaciones estables a nivel temporal y geográfico en ambientes homogéneos y controlados, con la excepción de las incertidumbres asociadas a la limitada representación de las condiciones que ocasionan una calidad muy mala del agua, lo que indujo a la obtención de predicciones sumamente variables, degradadas y carentes de un sentido físico o químico defendible; adicionalmente, el modelo falla rotundamente cuando las condiciones ambientales varían ligeramente, lo que demuestra la baja robustez en la estructura general de este, limitado por los parámetros iniciales de entrenamiento.

13

Validación del Modelo MARS.

La evaluación del modelo MARS permitió identificar un claro fallo en las estimaciones al aplicarse en condiciones heterogéneas; no obstante, el desempeño del ajuste mejoró notablemente a medida que la cobertura superficial del lago se observaba de manera más clara, sin la interferencia de nubes. En este orden de ideas, se evidenció una gran incertidumbre para determinar si el modelo generaba predicciones confiables; por lo tanto, se ideó una arquitectura de validación basada en el análisis de los residuos provenientes de la muestra de prueba reservada inicialmente, antes del ajuste general del MARS. Dicha muestra contiene un total de 212 observaciones de las 704 ubicaciones originales, es decir, el 30 % de los puntos en donde se calculó el índice a partir de los modelos de regresión lineal de los parámetros de calidad del agua.

La estructura general de la validación se resume en 3 fases esenciales: en la primera, se calcularon estadísticas de tendencia central y de conformidad en el ajuste, que ayudaron a analizar el alcance y la captura de la estructura espectral del índice; en la segunda, se crearon mapas de distribución de los residuos, los cuales permitieron identificar si existían patrones espaciales en los residuos, apoyados por pruebas de autocorrelación espacial para determinar si los residuos estaban fuertemente influenciados por la ubicación en el lago; finalmente, se ejecutó una prueba de significancia estadística por medio de una prueba F.

13.1. Métricas de Concordancia.

En el capítulo anterior, se calcularon diferentes métricas de desempeño para verificar cuál modelo presentaba mejor ajuste dentro de las cuatro muestras asociadas a cada estación del año. Dichas métricas permitieron concluir que ningún ajuste logró explicar la variabilidad del índice, dado que el mejor modelo presentó una suma de cuadrados menor que la de los residuos; por lo tanto, el coeficiente de determinación R^2 se desplomó; sin embargo, existen otras dos métricas que permiten analizar el grado de concordancia de las estimaciones; en otras palabras, la calidad del ajuste. Estas medidas son el *Index of Agreement* (IoA) y el *Scatter Index*.

13.1.1. Índice de Concordancia (*Index of Agreement*).

El IoA mide el grado de concordancia general entre los valores estimados y los observados; es decir, qué tan bien un modelo reproduce los patrones existentes (Duveiller et al., 2016). La estructura del índice se presenta en la ecuación (13.1), la cual genera valores en un rango de 0 a 1, siendo el límite máximo el indicador de concordancia perfecta; por otra parte, valores más cercanos a 0 indican que el índice fue afectado por sesgo y un grado muy alto de dispersión; por consiguiente, producen grandes errores.

$$IoA = \frac{\sum(P_i - O_i)^2}{\sum(|P_i - \overline{O}| + |O_i - \overline{O}|)^2} \tag{13.1}$$

Donde P_i es el valor estimado del índice, O_i es el valor observado del índice y $\overline{O}$ es la media observada de la variable dependiente. La interpretación típica del IoA se resume de la siguiente manera:

- IoA >0,7; buena concordancia.
- 0,5 >IoA >0,7; concordancia moderada.
- IoA <0,5; concordancia pobre.

13.1.2. Índice de Dispersión (*Scatter Index*).

El SI cuantifica la dispersión relativa de los errores con respecto a la magnitud de los datos; es decir, permite identificar si los errores son pequeños al relacionarlos con la magnitud de la variable independiente (Arias-Rodríguez et al., 2021). La ecuación (13.2) define la estructura del índice. A diferencia del IoA, este parámetro no tiene un rango establecido; por lo tanto, se busca obtener valores bajos, ya que cuanto más pequeña es la dispersión de los datos con respecto a la media, más bajo es el resultado de la división. De la misma forma, al obtenerse una dispersión muy alta con respecto al promedio, el índice tenderá a arrojar valores mucho más altos.

$$SI = \frac{RMSE}{\overline{O}} \tag{13.2}$$

Donde $RMSE$ es la raíz del error cuadrático medio y $\overline{O}$ es el promedio de las observaciones. La interpretación general del SI se resume de la siguiente forma:

- SI <0,3; buen desempeño, errores pequeños.
- 0,3 <SI <0,5; desempeño aceptable, errores moderados.
- SI >0,5; desempeño pobre, errores proporcionalmente grandes.

13.2. Estadísticas de los Residuos.

13.2.1. Estadísticas descriptivas.

La muestra elegida para los residuos tuvo un total de 212 observaciones, como ya se había mencionado anteriormente. El promedio fue de 0,2688, lo que indica que los residuos estuvieron concentrados muy cerca de 0, con una ligera tendencia a la subestimación determinada por el resultado positivo; no obstante, al identificar el mínimo, el máximo y los cuantiles, se evidenció una anomalía relacionada con el límite inferior, pues hubo una diferencia de casi 14 unidades con la estimación asociada. Esto quiere decir que el modelo no logró capturar la variabilidad de esta observación y que la predicción subsecuente estuvo muy por encima de lo esperado. A partir de este dato, se pudo inferir por qué el modelo mostró una tendencia a subestimar el índice cuando los valores espectrales eran muy altos.

Aun así, las demás estadísticas no presentaron un comportamiento sumamente disperso; por ejemplo, los cuantiles se mantuvieron por encima de 0, con una distancia moderada entre sí, llegando a un máximo de 2,835 y una desviación estándar de 2,75, bastante similar al error medio cuadrático obtenido en el ajuste del modelo. En síntesis, el modelo reflejó un sesgo ligeramente positivo; es decir, tiende a elevar el índice, que a su vez se castiga severamente cuando los niveles espectrales aumentan su valor; en consecuencia, los valores negativos son tan abultados. El resumen de las estadísticas descriptivas se muestra en la tabla (13.1).

TABLA 13.1: Resumen de las estadísticas descriptivas de los residuos.

Parámetro	**Medida**
Cantidad	212
Media	0,2688
Desv. Estándar	2,75
Mínimo	-13,98
25 %	0,424
50 %	1,256
75 %	1,711
Máximo	2,835

Fuente: Elaboración propia.

13.2.2. Estadísticas de validación.

La raíz del error cuadrático medio, junto con el error absoluto, ya se había presentado en el capítulo de evaluación del modelo, determinando que las métricas son razonables y no muestran un indicio fuerte de dispersión desmedida. Por lo tanto, se calculó el IoA y el SI, los cuales reflejaron valores de 0,224 y 0,033, respectivamente. Con respecto a la concordancia, se evidenció que el modelo arrojó un resultado pobre debido a su lejanía al limite superior del rango del índice. Esto indicaría que el modelo no reproduce correctamente la variabilidad del WQI; no obstante, el IoA tiene una limitación asociada a la baja variabilidad de la variable respuesta, en la medida en que cualquier residuo excesivo, ya sea superior o inferior, castiga severamente la métrica. Como ya se explicó en el capítulo anterior, la homogeneidad en la calidad del agua ocasionó que la dispersión en los datos fuera controlada; en consecuencia, tanto el coeficiente de determinación como el índice de concordancia presentaron valores notablemente pequeños.

Por otra parte, confirmando los hallazgos del IoA, el valor del SI fue significativamente bajo, por lo que se confirmó la baja variabilidad que afectó el cálculo de concordancia. En resumen, las estimaciones están concentradas alrededor de la media por falta de variabilidad, lo cual genera errores sistemáticos cuya estructura no se logró capturar espectralmente, ocasionando que el modelo sobrestime el índice, lo que a su vez forzó una penalización materializada en estimadores negativos muy susceptibles a valores espectrales elevados. Ante este panorama, se concluyó que el modelo tiene utilidad de manera limitada para estimar una clasificación general del WQI, pero carece de confiabilidad para realizar predicciones precisas en escenarios heterogéneos o en presencia de concentraciones prominentes en los parámetros físico-químicos que realmente estén afectando la calidad del agua. El resumen de las estadísticas de validación se muestra en la tabla (13.2).

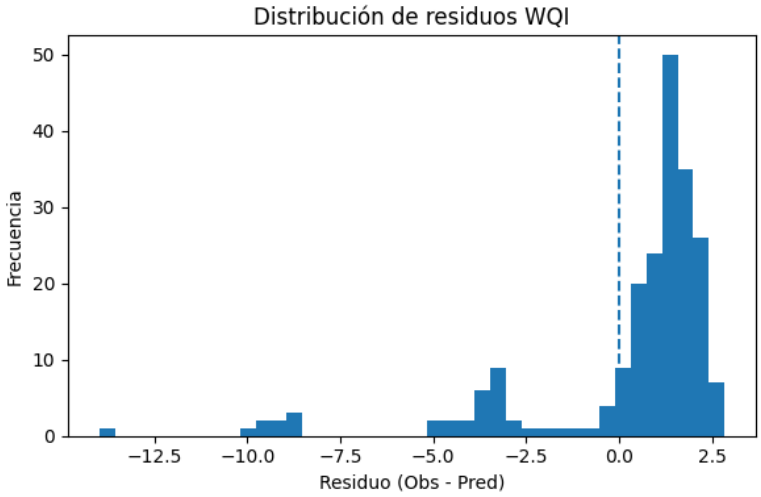


FIGURA 13.1: Histograma de los residuos del modelo MARS.
Fuente: Elaboración propia.

TABLA 13.2: Resumen de las estadísticas de validación de los residuos.

Parámetro	Medida
RMSE	2,760
MAE	1,989
IoA	0,224
SI	0,033

Fuente: Elaboración propia.

13.3. Distribución de los Residuos.

Para comprender mejor el comportamiento tan disparejo de los residuos, se obtuvo su histograma, el cual se presenta en la figura (13.1). A partir de esta gráfica, se puede identificar que el modelo está sólidamente forzado por una cola a la izquierda, lo que genera una concentración masiva de residuos a la derecha. En términos prácticos, los valores más bajos del índice se inflaron súbitamente; es decir, hacia la media; por lo tanto, los valores más altos sufrieron ligeras subestimaciones para controlarlos y equilibrarlos alrededor del valor central mencionado.

Con base en lo anterior, fue posible explicar por qué el modelo presentó valores aceptables en la raíz del error cuadrático medio, absoluto y relativo, junto con el índice de dispersión SI; y, a su vez, fue incapaz de explicar la variabilidad del WQI. Pues bien, evocando la fase de entrenamiento, el ajuste del MARS se realizó con una muestra cuya mayor concentración estaba en el rango relativo a la categoría buena del índice, dejando una pequeña cantidad de observaciones para la calidad aceptable y un solo registro para la calidad mala; ante este panorama, el modelo simplemente decidió ignorar aquellas perturbaciones que no representaban el estado general del lago; es decir, los valores del WQI menores que 80.

La figura (13.2) representa la distribución de los residuos con su correspondiente valor estimado. Nótese cómo el valor ya identificado como crítico (residuo de -14 puntos) obtuvo una estimación

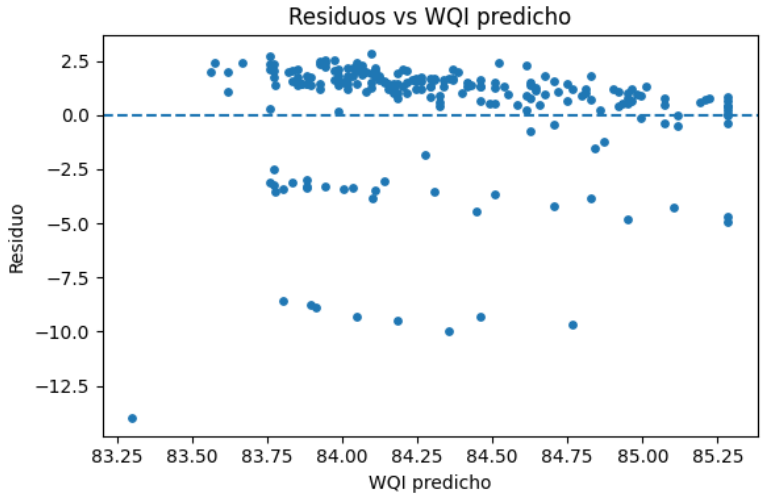


FIGURA 13.2: Distribución de los residuos del modelo MARS.
Fuente: Elaboración propia.

de aproximadamente 83,30, cuando en realidad su valor observado era de 69,32. Paralelamente, las siguientes observaciones tuvieron estimaciones entre 83,75 y 85,00, en contraste con el rango entre 74,39 y 75,15 de las observaciones asociadas. Estos nueve valores diferentes explicaron la estructura que el modelo realmente capturó, la cual no tomó en cuenta las condiciones que estaban generando una caída en la calidad del agua, sino que buscaba atraer todos los valores hacia la concentración masiva en la cual se halló la mayor cantidad de datos; es decir, el promedio.

En consecuencia, los residuos de la mencionada muestra reducida asociada a las zonas con calidad de agua más degradadas tuvieron valores excesivamente alejados de su valor observado, lo que confirmó la poca capacidad del modelo para capturar la dinámica real de la calidad del agua y, en su lugar, se limitó a estimar los valores de índice cuya combinación de bandas fuera la más parecida a la de la fase de entrenamiento. Por otra parte, el modelo castigó las predicciones para las combinaciones de bandas que fueran distintas en comparación con las de la fase de entrenamiento o que sencillamente presentaran lecturas excesivas en la banda 9, provocando una disminución físicamente inverosímil en el índice, indicando que la calidad del agua estaba afectada cuando, de hecho, no tenía la capacidad de predicción necesaria para generar dicha afirmación.

13.3.1. Distribución espacial.

La partición aleatoria de los datos fue satisfactoria en términos espaciales, como se muestra en la figura (13.3); puesto que la muestra elegida para validar el modelo no se consolidó en una parte específica del lago, sino que se seleccionaron ubicaciones a lo largo del conjunto válido de observaciones con las cuales se ajustaron los modelos para cada estación. Esta aclaración cobra gran relevancia, teniendo en cuenta el descubrimiento de la estructura capturada por la regresión, en el sentido de que la presencia de residuos tan altos podría haberse justificado por la incidencia de factores espaciales específicos y no por un fallo sistemático del ajuste.

No obstante, la espacialización de los residuos sugiere que no existe un patrón geográfico particular; es decir, las nueve observaciones con calidad aceptable que fueron estimadas como buenas se encuentran distribuidas a lo largo del lago; por lo tanto, no se encontró evidencia para afirmar que

el modelo falló en zonas particulares de la zona de estudio; por el contrario, la posible degradación representada en un índice menor no pudo ser explicada por el MARS mediante los valores de reflectancia de las bandas 3 y 9; por consiguiente, se desestimaron y se consideraron como localizaciones sin perturbaciones o afectaciones mayores.

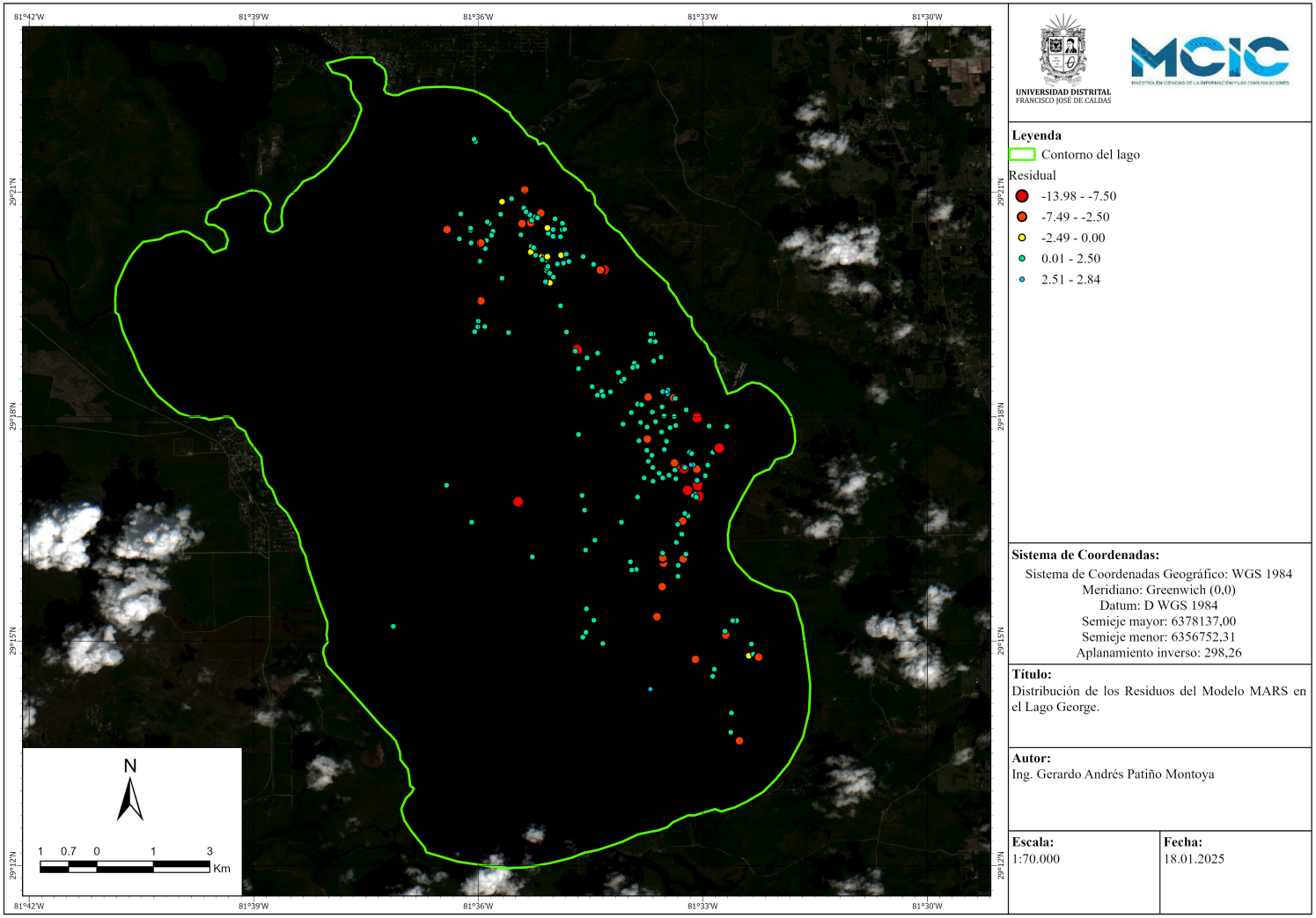


FIGURA 13.3: Distribución de los residuos del modelo MARS en el lago George.
Fuente: Elaboración propia a partir de imágenes Sentinel-2 (Copernicus, 2023).

13.4. Prueba de Autocorrelación Espacial.

La distribución espacial observada permitió identificar que los residuos no parecían tener algún tipo de correlación o tendencia geográfica; aún así, se aplicó la prueba I de Moran exclusivamente en el conjunto de datos de validación con el fin de confirmar los hallazgos demostrados en la figura (13.3). Dicha prueba es una medida ampliamente utilizada para la evaluación del nivel de correlación de los valores de una variable en un espacio, determinando si un conjunto de datos geográficos presenta un patrón agrupado, disperso o aleatorio, con base en sus ubicaciones y valores asociados (Moran, 1950). La interpretación del índice se explica a continuación:

- I = 0, *p*-valor >0,05; los residuos no presentan autocorrelación espacial.
- I >0, *p*-valor <0,05; los residuos están autocorrelacionados, estructura no capturada por el modelo.
- I <0, *p*-valor <0,05; dispersión espacial.

La hipótesis nula de la prueba indica que los residuos están distribuidos aleatoriamente; mientras que la hipótesis alterna rechaza la aleatoriedad, apuntando a una estructura espacial no explicada por el modelo. Los resultados del análisis no mostraron evidencia de dependencia espacial significativa en los residuos, respaldados por un valor del I de Moran de 0,0026 y una probabilidad asociada mayor al nivel de significancia (0,05), como lo muestra la tabla (13.3). En ese orden de ideas, se pudo confirmar la hipótesis de que los residuos del modelo MARS se distribuyen aleatoriamente a lo largo de la extensión del lago George, sin dejar patrones geográficos sistemáticos sin capturar.

TABLA 13.3: Resultados de la prueba I de Moran para la detección de autocorrelación espacial.

Parámetro	**Valor**
I de Moran	0,0026
E[I]	0,0047
z-score	0,234
p-valor (normal)	0,815
p-valor (permutaciones)	0,374

Fuente: Elaboración propia.

13.5. Bootstrap y Validación Cruzada.

Con el propósito de evaluar la estabilidad y la incertidumbre de las métricas de desempeño del modelo, se implementó un procedimiento de remuestreo *Bootstrap*. Esta técnica consiste en generar múltiples muestras del mismo tamaño que el conjunto de datos original mediante muestreo con reemplazo, de modo que algunas observaciones pueden repetirse mientras otras pueden quedar excluidas en cada iteración. A partir de estas muestras se reentrena y evalúa el modelo, obteniendo una distribución empírica de las métricas de desempeño. Este procedimiento permite cuantificar la sensibilidad del modelo frente a variaciones en los datos de entrenamiento y estimar la variabilidad asociada a sus indicadores de desempeño, sin requerir supuestos paramétricos sobre la distribución de los errores.

El procedimiento se implementó mediante 500 iteraciones de remuestreo con reemplazo, generando en cada réplica una muestra del mismo tamaño que el conjunto de datos original, como ya se

mencionó anteriormente. Para cada iteración se entrenó nuevamente el modelo MARS utilizando la misma configuración de hiperparámetros (max_degree = 2 y penalty = 3.0), con el fin de que la única fuente de variabilidad correspondiera a la composición de la muestra de entrenamiento. Posteriormente, el desempeño del modelo se evaluó utilizando únicamente las observaciones *Out-of-Bag*, es decir, aquellas no seleccionadas durante el remuestreo, proporcionando así una estimación independiente del error de generalización. Finalmente, se resumió la distribución empírica de las métricas RMSE, MAE y R^2 mediante su media, desviación estándar y los percentiles 2,5 % y 97,5 %, los cuales constituyen un intervalo empírico del 95 %.

Los resultados del análisis evidenciaron una elevada variabilidad del RMSE entre las diferentes muestras, reflejada tanto en su desviación estándar (23,78) como en el amplio intervalo comprendido entre los percentiles 2,5 % y 97,5 %. Este comportamiento indica que el desempeño predictivo del modelo dependió considerablemente de las observaciones seleccionadas durante el entrenamiento, por lo que pequeñas modificaciones en la composición de la muestra produjeron cambios importantes en el error de predicción. En consecuencia, el modelo presentó una estabilidad limitada y una capacidad reducida de generalización.

En contraste, el MAE presentó una variabilidad considerablemente menor, indicando que la mayor parte de las predicciones generaron errores moderados. La discrepancia entre la estabilidad del MAE y la elevada dispersión del RMSE sugiere la presencia de errores puntuales de gran magnitud, los cuales incrementaron de forma desproporcionada el RMSE debido a la naturaleza cuadrática de esta métrica. Los valores negativos del coeficiente de determinación, obtenidos en la mayoría de las iteraciones, indican que el modelo presentó un desempeño inferior al de un predictor menos robusto basado en la media de la variable respuesta. Además, la elevada dispersión observada confirma que la capacidad explicativa del modelo fue altamente inestable frente a pequeñas variaciones del conjunto de entrenamiento.

TABLA 13.4: Resultados de la *Bootstrap* para el conjunto de datos completo.

Parámetro	**Media**	**Desv. Estándar**	**2,5 %**	**97,5 %**
RMSE	7,197	23,776	2,650	28,328
MAE	2,415	1,488	1,958	3,855
R^2	-80,285	1549,156	-100,045	0,021

Fuente: Elaboración propia.

Por otra parte, se realizó una rutina de validación cruzada tipo *k-fold*, la cual divide el conjunto de datos en subconjuntos; luego, los utiliza para entrenar el modelo, dejando uno para la validación, repitiendo el proceso tantas veces como subconjuntos se hayan dividido; finalmente, se promedian las medidas de desempeño y se analiza la capacidad de generalización del modelo general (Pérez-Planells et al., 2015). Los resultados del procedimiento explicado anteriormente no indicaron una inestabilidad excesiva en cada uno de los grupos seleccionados, evidenciando una variabilidad del RMSE más moderada y alineada con las medidas obtenidas en el ajuste inicial. Este mismo patrón se presentó en el error medio absoluto, el cual no sufrió cambios súbitos entre las particiones. Finalmente, se confirmó nuevamente la escasa capacidad del modelo para explicar la variabilidad del índice, lo que se traduce en predicciones poco confiables. El resumen de la validación cruzada se muestra en la tabla (13.5).

TABLA 13.5: Resultados de la validación cruzada.

Parámetro	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
RMSE	2,996	2,628	3,909	2,557	2,822
MAE	2,071	1,994	2,470	1,912	2,109
R^2	0,002	0,031	-0,173	-0,088	0,027

Fuente: Elaboración propia.

13.6. Prueba de Significancia.

La validación final del modelo se realizó verificando la significancia estadística del MARS con los datos de entrenamiento mediante la prueba F, la cual define si la hipótesis de que el valor de la variación calculada con respecto al modelo de regresión es mayor que el valor estimado con base en promedios es válida; en otras palabras, si el modelo predice mejor el índice en comparación con su media aritmética (Berger et al., 2018).

Para tal fin, se obtuvo la suma de cuadrados del modelo o SSE en la ecuación (13.3), la suma de cuadrados de los residuos de las observaciones de entrenamiento o SSR en la ecuación (13.4) y el error cuadrático medio tanto del modelo o MSE en la ecuación (13.5), como de los residuos o MSR en la ecuación (13.6). El grado de significancia α fue del 5 % y los grados de libertad (λ en la ecuación (13.7)) se estimaron con base en el número total de registros (492) menos el número de variables independientes, que para este caso se usaron 11 bandas, menos uno; para un total de 480.

$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{13.3}$$

$$SSR = \sum (\hat{y}_i - \overline{y})^2 \tag{13.4}$$

$$MSE = \frac{SSE}{\lambda} \tag{13.5}$$

$$MSR = \frac{SSR}{\gamma - 1} \tag{13.6}$$

$$\lambda = U - \gamma - 1 \tag{13.7}$$

Una vez que se calculó el MSE y MSR, se reemplazan en la ecuación (13.8) y se obtiene el valor práctico del estadístico F para el modelo MARS. Por otra parte, la prueba de hipótesis se muestra en la ecuación (13.9), de tal forma que, si el valor práctico es mayor que el teórico, se rechaza la hipótesis nula por falta de evidencia de que el modelo no se ajusta a los datos a un nivel de significancia del 5 %.

$$F_0 = \frac{MSR}{MSE} \tag{13.8}$$

$$H_o : \beta_1, \beta_2, ..., \beta_i = 0$$
$$H_a : \beta_1, \beta_2, ..., \beta_i \neq 0 \tag{13.9}$$

Los resultados de los estadísticos para la prueba de significancia del modelo MARS se presentan en la tabla (13.6). El estadístico F_0, con base en los errores medios cuadráticos de la muestra de entrenamiento, arrojó un valor de 6,12. Paralelamente, se determinó el valor teórico con base en los parámetros mencionados anteriormente ($F_{0,05;10,480}$), cuyo valor fue de 1,84; por lo tanto, como el F_0 >Fcrítico, se rechaza la hipótesis nula y se concluye que el modelo es estadísticamente significativo.

TABLA 13.6: Resultados de la prueba F para la significancia estadística del modelo.

Estadístico	**Valor**
SSR	509,93
SSE	3997,82
MSR	50,99
MSE	8,33
F_0	6,12

Fuente: Elaboración propia.

Finalmente, la rutina de validación creada permitió establecer que, si bien el modelo era estadísticamente significativo y presentaba errores relativamente bajos, su capacidad para capturar la variabilidad y la estructura general del índice de calidad del agua en el lago George fue baja; por lo tanto, las estimaciones obtenidas no pudieron considerarse completamente precisas. Lo anterior no significó que el modelo no fuera útil; por el contrario, su implementación y posterior evaluación podrían contribuir a la identificación de perturbaciones en la calidad del agua cuando las predicciones no se encuentren dentro del rango de la media general, que es predominantemente buena.

14

Conclusiones.

La presente investigación permitió desarrollar un modelo que evaluara el índice de calidad del agua del CCME en el lago George mediante la integración de información proveniente de imágenes satelitales del programa Sentinel-2 con técnicas de aprendizaje automático, representando una alternativa realizable con el fin de apoyar el monitoreo de la calidad del agua en el cuerpo de agua mencionado anteriormente; no obstante, su desempeño y significancia estadística asociada estuvieron sistemáticamente relacionados con la calidad, frecuencia y representatividad de las mediciones de campo que sirvieron de insumo para los modelos de regresión lineal. En términos generales, se evidenció que el enfoque tradicional basado en ecuaciones lineales resultó notablemente limitado para capturar la estructura espectral y temporal de los parámetros elegidos, debido a la naturaleza estacional propia de las muestras, la cual presentó una dificultad significativa en cuanto al modelado lineal de los datos. Por otra parte, la metodología del modelo MARS produjo menores errores; no obstante, presentó un pobre ajuste general, afectado significativamente por una homogeneidad generalizada del CCME WQI y por perturbaciones espectrales asociadas a la presencia de nubes en las imágenes utilizadas. Por lo tanto, los resultados derivados de la evaluación y validación del modelo demostraron que las técnicas de teledetección pueden complementar los esquemas convencionales de seguimiento y monitoreo ambiental, pero deben acompañarse con planes de mediciones más robustos y temporalmente consistentes con el objetivo de diagramar de forma más precisa la naturaleza no lineal y dinámica tanto de los indicadores de calidad del agua como de la integridad propia del cuerpo de agua en cuestión.

La implementación de una arquitectura basada en modelos de regresión lineal multivariada, con el fin de diagramar el comportamiento espectral de los parámetros de calidad del agua en el lago George, presentó varios matices, desde la limpieza de datos hasta la estimación y validación de las predicciones generadas a partir de las mismas ecuaciones del ajuste. Inicialmente, fue imperativa la necesidad de eliminar una gran porción de observaciones para lograr un coeficiente de determinación ajustado aceptable, lo cual provocó que la robustez de muchos modelos se redujera a una muestra notablemente limitada. Esto, a su vez, generó que las concentraciones estimadas fueran muy susceptibles a valores extremos en los valores de reflectancia de las bandas involucradas en las ecuaciones correspondientes, como fue el caso de la clorofila-a, cuyo modelo estimó valores excesivamente altos y carentes de algún sentido físico o químico lógico. Por lo tanto, el enfoque original, sustentado en ecuaciones lineales, podría no haber sido el adecuado en la medida en que las muestras finales, con las que se realizaron los ajustes, llegaron a contener menos de la mitad de los registros que, preliminarmente, se habían adquirido. Adicionalmente, al tratarse de mediciones realizadas en un periodo considerablemente extenso, la influencia del componente temporal en los modelos ajustados mediante mínimos cuadrados ordinarios fue significativamente notable, hasta el punto de forzar un cambio de metodología, dado que la predominante autocorrelación positiva en los residuos impedía que tanto las estimaciones como la validación del modelo fueran confiables. En consecuencia, la

implementación de los mínimos cuadrados generalizados con autocorrelación permitió neutralizar dicha influencia temporal en los residuos, sin comprometer de manera abrupta el ajuste general de los análisis ni la verificación de los demás supuestos clásicos de la regresión lineal; no obstante, para variables como el nitrógeno, el fósforo y el potasio, no fue posible corregir la correlación temporal mencionada anteriormente, lo que significó su directa desestimación en las fases posteriores de la presente investigación. El oxígeno disuelto y la profundidad del disco Secchi también fueron eliminados del análisis, pero por motivos diferentes: para el primero, no se encontró relación lineal entre las mediciones y los valores de reflectancia de las bandas sino hasta que se removió más del 75 % de la muestra original; para el segundo, la detección de heterocedasticidad estructural impidió que los errores estándar y los intervalos de confianza asociados de los estimadores fueran confiables.

La generación del modelo MARS se realizó tras la estimación de los parámetros a partir de las ecuaciones de regresión lineal validadas, mediante el calculo del índice correspondiente con base en la metodología del Consejo Canadiense de Ministros del Medio Ambiente, cuyos resultados fueron principalmente homogéneos en el sentido de que cerca del 95 % de la muestra total evaluable presentó un índice en el rango de 80 a 85, relacionado con una buena calidad del agua. El 5 % restante se constituyó por registros del índice en las categorías aceptable y una sola observación categorizada como mala. Dicha homogeneidad provocó que el modelo MARS no fuera capaz de capturar la completa variabilidad del índice en el lago, agrupando la mayor parte de las estimaciones alrededor de la media de la muestra con la que inicialmente se entrenó. Ante este panorama, se obtuvieron coeficientes de determinación extremadamente bajos para cada uno de los cuatro modelos entrenados, dejando el de la imagen del mes de julio como el que mejor se ajustó, con un R^2 de 0,009. Si bien el enfoque del modelo MARS se centró en los valores predominantes de la muestra de entrenamiento, la raíz del error cuadrático medio, así como los errores absolutos y relativos, indicaron que el desempeño del mejor modelo no estuvo totalmente erróneo en la medida en que el grado de dispersión de los residuos fue considerablemente controlado, al obtenerse un promedio de 0,27 en la muestra de validación y una desviación estándar de 2,75. Esto significa que el modelo tendió a subestimar ligeramente las predicciones para equilibrar los valores máximos generados por las observaciones, con una calidad de agua ligeramente más degradada, cuyo índice fue elevado por un modelo fuertemente influenciado por la tendencia anteriormente mencionada.

La evaluación del CCME WQI a partir del modelo obtenido con base en la respuesta espectral de la imagen del mes de julio en la zona de estudio permitió identificar una fuerte influencia negativa del ajuste en presencia de coberturas nubosas, justificada por el hecho de que dos funciones bisagra activaban los estimadores asociados que degradaban el índice de manera desproporcionada; es decir, cuando la banda 9, que captura vapor de agua, leía valores de reflectancia altos, el modelo interpretaba dicha actividad espectral como perturbaciones nocivas en la superficie del lago, lo que se tradujo en la activación de estimadores negativos proporcionalmente grandes que, a su vez, provocaban que las predicciones presentaran valores muy bajos. Este fenómeno se presentó en las tres imágenes dispuestas para la evaluación del modelo; sin embargo, en las dos primeras, correspondientes a las imágenes de julio y agosto, se observó una mayor falencia en el desempeño general del modelo, al presentarse una cobertura de nubes prácticamente en la mitad de la extensión del cuerpo de agua objeto de análisis de la presente investigación. Por el contrario, la fase de evaluación en la tercera imagen correspondiente al mes de septiembre arrojó mejores resultados en la medida en que las estimaciones estaban más alineadas con los datos de entrenamiento en comparación con las imágenes de los dos meses inmediatamente anteriores; ya que cerca del 96 % de las predicciones indicaron que la calidad del agua en estos puntos era buena; mientras que esta misma consideración en los meses de julio y agosto difícilmente fue del 50 %, otorgando una mayor proporción de las estimaciones para

las peores categorías del índice; es decir, aceptable, mala y muy mala, estando esta última categoría completamente ausente en el cálculo inicial, lo que demuestra la poca capacidad del modelo para generar predicciones confiables y físicamente verosímiles ante condiciones gravemente afectadas por la presencia de nubes.

La validación del mejor modelo se realizó mediante el análisis del comportamiento de los residuos. En esta fase se evidenció la estructura que el modelo realmente capturó, la cual desestimó totalmente aquellas porciones de registros que no presentaban una calidad de agua categorizada como buena; por lo tanto, cualquier valor fuera del rango entre 80 y 85 se veía súbitamente aumentado, con el fin de alinearlo con la muestra mayor. Ante este panorama, en la muestra de validación se hallaron residuos irrisoriamente grandes, lo que a su vez provocó que la suma de cuadrados se inflara, desplomando el ajuste general del modelo. Por medio de la prueba de autocorrelación espacial se confirmó que los residuos no tenían patrones geográficos definidos, por lo que la obtención de estos no se debió a un fallo en la captura del esquema espacial del índice, sino a una falla inherente de variabilidad que imposibilitó mejorar las predicciones; en otras palabras, el modelo consideró que si el 95 % de los puntos evaluados presentan una buena calidad del agua, es probable que esta misma tendencia se replique en todo el lago, ignorando cualquier tipo de perturbación o afectación tanto espectral como ambiental que pudiera tener incidencia directa en la integridad físico-química del cuerpo lacustre. En síntesis, el modelo no logró capturar la estructura espectral del índice debido a su baja variabilidad, basada en las estimaciones de los modelos de regresión lineal asociados con los parámetros de calidad del agua, cuyas mediciones fueron adquiridas en el periodo comprendido entre septiembre de 2022 y abril de 2024, lo que ocasionó que la suma de cuadrados de los residuos fuera más grande que la del modelo, resultando en un coeficiente de determinación excesivamente bajo; aún así, el desempeño y la significancia general del ajuste no fueron del todo pobres, dado que los errores en la muestra de validación produjeron valores considerablemente bajos y que la prueba F no halló evidencia de que el modelo no se ajustara a los datos.

15

Principales Aportes.

La elección del modelo de splines adaptativos multivariados (MARS) como técnica de aprendizaje automático para relacionar espacialmente el índice de calidad del agua con la respuesta espectral del lago George no respondió únicamente a consideraciones de desempeño estadístico frente a otras alternativas, sino también a una decisión deliberada de explorar metodologías de machine learning distintas a las predominantes en la literatura consultada. En efecto, la revisión bibliográfica realizada en el marco de esta investigación evidenció una concentración marcada de trabajos previos en teledetección de calidad del agua hacia el uso de redes neuronales artificiales, en sus distintas variantes (perceptrón multicapa, redes convolucionales, arquitecturas híbridas), como técnica preferente para la estimación de parámetros fisicoquímicos a partir de imágenes satelitales. Si bien dichas arquitecturas han demostrado buenos resultados en múltiples contextos, su naturaleza de caja negra limita la interpretabilidad de las relaciones entre bandas espectrales y parámetros estimados. Ante este panorama, se optó de manera intencional por un modelo de naturaleza no lineal pero interpretable, cuyas funciones bisagra permiten identificar explícitamente los puntos de corte y las combinaciones de bandas que determinan el comportamiento del índice, con el fin de contribuir a diversificar el repertorio metodológico documentado en la literatura regional sobre el tema, más allá de la reproducción de enfoques ya ampliamente explorados.

El aporte principal de la presente investigación no radica en la magnitud de los coeficientes de ajuste obtenidos, sino en la aplicación de una metodología alternativa de estimación de calidad del agua mediante teledetección en un cuerpo de agua que cuenta con una base de datos de monitoreo comparativamente robusta (en términos de frecuencia y cobertura temporal de mediciones fisicoquímicas), pero que se ubica en una región geográfica con escasa producción académica y bibliográfica previa en el uso de sensores remotos para el seguimiento de la calidad del agua. A diferencia de los antecedentes metodológicos más cercanos, desarrollados sobre cuerpos de agua colombianos previamente estudiados en el contexto de la teledetección nacional, el lago George representa un caso de aplicación en un territorio donde la integración de datos de monitoreo ambiental con imágenes Sentinel-2 no había sido explorada previamente, lo que constituye en sí mismo un ejercicio de validación metodológica en un contexto novedoso, más que una optimización incremental sobre un caso de estudio ya consolidado en la literatura.

En este sentido, el valor del presente trabajo debe entenderse como la evidencia de que es posible implementar una arquitectura metodológica completa (desde la limpieza de datos y la selección de variables espectrales, pasando por el ajuste y validación de modelos de regresión lineal para parámetros individuales, hasta la estimación espacial de un índice integrado mediante un modelo no lineal interpretable) en un cuerpo de agua con características de disponibilidad de datos distintas a las de los antecedentes nacionales, documentando de manera transparente tanto sus alcances como sus limitaciones.

La metodología desarrollada en la presente investigación se plantea como una herramienta exploratoria potencialmente aplicable a cuerpos de agua del territorio colombiano, bajo la condición indispensable de contar con una base de datos de mediciones de parámetros de calidad del agua sólida, frecuente y temporalmente consistente con la disponibilidad de imágenes Sentinel-2. Los resultados obtenidos evidenciaron que la robustez de los modelos de regresión lineal depende directamente de la cantidad, calidad y periodicidad de las observaciones de campo disponibles para el entrenamiento; en consecuencia, la replicabilidad de esta metodología en cuerpos de agua colombianos (muchos de los cuales enfrentan limitaciones de monitoreo sistemático) no debe entenderse como una aplicación directa e inmediata, sino como una hoja de ruta metodológica condicionada a un fortalecimiento previo de las redes de monitoreo ambiental local.

En aquellos casos donde dicha condición se cumpla, la presente investigación ofrece una arquitectura de trabajo replicable: la identificación de variables espectrales significativas mediante selección progresiva y control de multicolinealidad, la validación de los supuestos clásicos de la regresión lineal con corrección de autocorrelación temporal, y la estimación espacial de un índice de calidad integrado mediante técnicas de aprendizaje automático interpretables. Se propone, en consecuencia, que futuros trabajos orientados a la aplicación de esta metodología en cuerpos de agua colombianos prioricen, como paso previo, un diagnóstico de la disponibilidad y calidad de las series históricas de monitoreo fisicoquímico, de manera que la implementación de modelos de teledetección parta de una base de datos capaz de sostener la robustez estadística que este tipo de enfoques exige.

16

Recomendaciones.

Uno de los desafíos más grandes que se presentaron en el desarrollo de la presente investigación estuvo relacionado con la implementación de una metodología basada en modelos de regresión lineal, con el fin de diagramar el comportamiento espectral de los parámetros de calidad del agua en el lago; por lo tanto, un cambio de enfoque más robusto, como modelos autorregresivos más complejos, podría generar resultados más satisfactorios sin la necesidad de eliminar una cantidad significativa de datos. Adicionalmente, debido a la clara influencia del componente temporal y geográfico, la utilización de procedimientos relacionados con series de tiempo o regresiones con pesos espaciales tendría un papel crucial, con el objetivo de incluir en el análisis a aquellas variables que no fueron significativas o que presentaron un ajuste pobre.

El índice calculado con base en las estimaciones de los parámetros se realizó teniendo en cuenta una sola escala temporal; es decir, no se distinguió una posible fluctuación en la calidad del agua debido a que no todos los indicadores lograron ser evaluados en las 20 imágenes disponibles, y según el CCME, se recomienda evaluar el índice teniendo al menos cuatro registros por variable. En ese orden de ideas, la adquisición y el procesamiento de más imágenes intermedias representarían una mejor cobertura de las predicciones, permitiendo añadir el componente temporal en el modelo MARS, de tal forma que se incluyan fenómenos climáticos y atmosféricos estacionales en el análisis, los cuales no fueron relacionados con las técnicas empleadas en el presente proyecto.

Si bien la metodología del Consejo Canadiense de Ministros del Medio Ambiente es sólida y ampliamente utilizada en el análisis de la calidad del agua, no es la única que existe; por ejemplo, la Fundación Nacional de Saneamiento (NSF) desarrolló un índice para evaluar el nivel de calidad de un cuerpo hídrico basado en la medición de parámetros físico-químicos. En consecuencia, la inclusión de índices alternativos ayudaría a corroborar o desmentir los resultados del cálculo realizado, los cuales indicaron que la calidad del agua en el lago era predominantemente buena.

El modelo MARS demostró ser efectivo en la captura de la estructura del índice en trabajos previamente consultados; sin embargo, también mostró una clara falencia ante variables con un grado de dispersión notablemente limitado, como lo evidenciaron los resultados expuestos anteriormente. Es así que la exploración de métodos alternativos ofrecería una perspectiva adicional, con el fin de obtener estimaciones del índice de una manera más confiable y precisa. Adicionalmente, dado que la arquitectura de la metodología incluye una categorización asociada con el valor del índice, la inclusión de modelos de clasificación también contribuiría a mejorar la comprensión del comportamiento de la calidad del agua en el lago George.

Bibliografía

Ahmed, M., Mumtaz, R., and Anwar, Z. (2022a). An enhanced water quality index for water quality monitoring using remote sensing and machine learning. *Applied Sciences*.

Ahmed, M., Mumtaz, R., Anwar, Z., Shaukat, A., Arif, O., and Shafait, F. (2022b). A multi–step approach for optically active and inactive water quality parameter estimation using deep learning and remote sensing. *Water*.

Ahmed, M., Mumtaz, R., Baig, S., and Hassan-Zaidi, S. M. (2022c). Assessment of correlation amongst physico-chemical, topographical, geological, lithological and soil type parameters for measuring water quality of rawal watershed using remote sensing. *Water Supply*.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*.

Arias-Rodríguez, L. F., Duan, Z., Díaz-Torres, J. J., Basilio-Hazas, M., Huang, J., Udhaya Kumar, B., Tuo, Y., and Disse, M. (2021). Integration of remote sensing and mexican water quality monitoring system using an extreme learning machine. *Sensors*.

Astrofsky, K.M. abd Bullis, R. and Sagerstrom, C. (2002). Chapter 19 - biology and management of the zebrafish. *Academic Press*.

Bendel, R. B. and Afifi, A. A. (1977). Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association*.

Berger, P. D., Maurer, R. E., and Celli, G. B. (2018). *Experimental Design*. Springer.

Boehmke, B. and Greenwell, B. (2020). *Hands-On Machine Learning with R*. CRC Press.

Buse, A. (1973). Goodness of fit in generalized least squares estimation. *The American Statistician, Vol. 27*.

CCME (2017). *CCME Water Quality Index User's Manual 2017 Update*. Canadian Council of Ministers of the Environment.

Chang, N., Xuan, Z., and Yang, Y. J. (2013). Exploring spatiotemporal patterns of phosphorus concentrations in a coastal bay with modis images and machine learning models. *Remote Sensing of Environment*.

Chebud, Y., Naja, G. M., Rivero, R., and Melesse, A. (2012). Water quality monitoring using remote sensing and and artificial neural network. *Water, Air and Soil*.

Chen, Y. (2016). Spatial autocorrelation approaches to testing residuals from least squares regression. *PLoS ONE*.

Clarke, J., Williams, L., and Cherry, G. (2010). Hydrogeology and water quality of the floridan aquifer system and effect of lower floridan aquifer pumping on the upper floridan aquifer at hunter army airfield, chatham county, georgia. Technical report, United States Geological Survey(USGS).

Copernicus (2023). Sentinel-2 satellite imagery. Copernicus Programme. Imágenes satelitales, nivel L2A.

Cumming, J. (2006). *A Brief Florida Real Estate History*. Appraisal Institute.

Duveiller, G., Fasbender, D., and Meroni, M. (2016). Revisiting the concept of a symmetric index of agreement for continuous datasets. *Scientific Reports*.

El Din, E. S., Zhang, Y., and Suliman, A. (2017). Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. *Remote Sensing*.

EPA (1976). Quality criteria for water. Technical report, United States Environmental Protection Agency. Disponible en: https://www.epa.gov/sites/default/files/2018-10/documents/quality-criteria-water-1976.pdf. Accedido el 22 de diciembre de 2025.

EPA (1986). Quality criteria for water. Technical report, United States Environmental Protection Agency. Disponible en: https://www.epa.gov/sites/default/files/2018-10/documents/quality-criteria-water-1986.pdf. Accedido el 22 de diciembre de 2025.

EPA (2006). National primary drinking water regulations. Technical report, United States Environmental Protection Agency. Disponible en: https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations. Accedido el 19 de diciembre de 2025.

EPA (2012). National lakes assessment 2012: A collaborative survey of lakes in the united states. Technical report, United States Environmental Protection Agency. Disponible en: https://www.epa.gov/sites/default/files/2016-12/documents/nla_report_dec_2016.pdf. Accedido el 21 de diciembre de 2025.

EPA (2023). Secondary drinking water standards: Guidance for nuisance chemicals. Technical report, United States Environmental Protection Agency. Disponible en: https://www.epa.gov/sdwa/secondary-drinking-water-standards-guidance-nuisance-chemicals. Accedido el 19 de diciembre de 2025.

ESA (2015). Sentinel 2. Technical report, European Spatial Agency.

ESA (2021). Copernicus sentinel-2. msi level-2a boa reflectance product. collection 1. Technical report, European Space Agency.

ESA (2026). S2 products. msi products overview. Technical report, European Space Agency.

Friedman, J. H. (1991). Multivariate adaptative regression splines. *The Annals of Statistics, JSTOR*.

Galton, F. (1886). *Family Likeness in Stature*. Proceeding of Royal Society.

Gholizadehm, M. and Melesse, A. (2017). Study on spatiotemporal variability of water quality parameters in florida bay using remote sensing. *Journal of Remote Sensing GIS*.

Google Maps (2026). Google maps. https://www.google.com/maps/place/Lake+George/@29.2864326,-81.5991094,31465m/data=!3m2!1e3!4b1!4m6!3m5!1s0x88e656d343389e09:0xa0584f045373be2d!8m2!3d29.2685061!4d-81.5815776!16zL20vMDN4NGpm?entry=ttu&g_ep=EgoyMDI2MDQwNy4wIKXMDSoASAFQAw%3D%3D. Captura de pantalla, consultado el 10 de abril de 2026.

Government of Canada (1992). Guidelines for canadian drinking water quality: Guideline technical document – sodium. Technical report, Government of Canada. Disponible en: https://www.canada.ca/en/health-canada/services/publications/healthy-living/guidelines-canadian-drinking-water-quality-guideline-technical-document-sodium.html. Accedido el 22 de diciembre de 2025.

Guerrero Buelbas, L. M. (2022). Modelo para la predicción de la calidad del agua utilizando imágenes sentinel a partir de las concentraciones físicoquímicas. caso de estudio lago de tota. Master's thesis, Universidad Distrital Francisco José de Caldas.

Gujarati, D. and Porter, D. (2010). *Econometría*. Mc Graw Hill.

Hart, W. (1998). Alternative formulations of factors used in the british columbia water quality index. Technical report, Report submitted to the Technical Subcommittee of the CCME Water Quality Guidelines Task Group.

Husain, T. (1998). Application of the national water quality index prototype in newfoundland. Technical report, Prepared for State of the Environment Task Group, Canadian Council of Ministers of the Environment.

Hébert, S. (1996). Développement d'un indice de la qualité bactériologique et physico-chimique de l'eau pour des rivières du québec. *Report of the Ministère de l'Environnement et de la Faune*.

INSST (2017). Limpieza y mantenimiento en espacios confinados de estaciones depuradoras de aguas residuales: exposición a sulfuro de hidrógeno. *Instituto Nacional de Seguridad y Salud en el Trabajo.*

Kang, I. and Kug, J. S. (2002). El niño and la niña sea surface temperature anomalies: Asymmetry characteristics associated with their wind stress anomalies. *Journal of Geophysical Research*.

LCWA (2025). Glossary. Technical report, Lake County Water Atlas. Disponible en: https://lake.wateratlas.usf.edu/library/glossary/. Accedido el 14 de mayo de 2025.

Lumumba, V., Kiprotich, D., Lemasulani, M., Grace, N., and M., D. (2024). Comparative analysis of cross-validation techniques: Loocv, k-folds cross-validation, and repeated k-folds cross-validation in machine learning models. *SSRN*.

Lund, D. C. and Curry, W. (2006). Florida current surface temperature and salinity variability during the last millennium. *Paleoceanography and Paleoclimatology*.

Madigan, M., Martinko, J., Stahl, D., and Clark, D. (2010). *Brock Biology of microorganisms*. Pearson Benjamin Cummings, UK, 13th edition.

Marella, R. and Berndt, M. (2005). Water withdrawals and trends from the floridan aquifer system in the southeastern united states, 1950-2000. *U.S. Geological Survey Circular 1278*.

Marin, R. (2024). Florida superó los 23 millones de habitantes y se consolidó como el tercer estado más poblado de estados unidos. *Infobae*.

Martínez, E. (2005). Errores frecuentes en la interpretación del coeficiente de determinación lineal. *Anuario Jurídico y Económico Escurialense.*

Morales-Oñate, V. (2025). Machine learning: Teoría y práctica. Technical report, Universidad de Las Américas.

Moran, P. (1950). Notes on continuous stochastic phenomena. *Oxford Journals*.

Najafzadeh, M. and Basirian, S. (2023). Evaluation of river water quality index using remote sensing and artificial intelligence models. *Remote Sensing*.

National Research Council (2012). Review of the st. johns river water supply impact study: Final report. Technical report, The National Academies Press.

Neary, B., Cash, K., Hébert, S., Khan, H., Saffran, K., Swain, L., Williamson, D., and Wright, R. (2001). Ccme water quality index 1.0 technical report. *Canadian Water Quality Guidelines for the Protection of Aquatic Life.*

NIH (2019). Datos sobre el potasio. Technical report, National Institutes of Health: Office of Dietary Supplements.

Pereira-Sandoval, M., Ruescas, A. Urrego, P., Ruiz-Verdú, A., Delegido, J., Tenjo, C., Soria-Perpinyá, X., Vicente, E., Soria, J., and Moreno, J. (2019). Evaluation of atmospheric correction algorithms over spanish inland waters for sentinel-2 multi spectral imagery data. *Remote Sensing.*

Phippen, B. (1998). Application of the national water quality index prototype in british columbia. *BWP Consulting*.

Piehler, M. F., Dyble, J., Moisander, P. H., Chapman, A., Hendrickson, J., and Paerl, H. (2009). Interactions between nitrogen dynamics and the phytoplankton community in lake george, florida, usa. *Lake and Reservoir Management*.

Porter, J. (2007). Lake george, information guide. *Florida Freshwater Fishing Series*.

Pérez, F. and Fernández, H. (2008). *Econometría: Conceptos Básicos*. ECOE Ediciones.

Pérez-Planells, L., Delegido, J., Rivera-Caicedo, L., and Verrelst, J. (2015). Analysis of cross-validation methods for robust retrieval of biophysical parameters. *Revista de Teledetección.*

Rocchini, R. and Swain, L. (1995). The british columbia water quality index. *Water Quality Branch, Environmental Protection*.

Salas-Salvadó, J., Maraver, F., Rodríguez-Mañas, L., Sáenz de Pipaon, M., Vitoria, I., and Moreno, L. (2021). Importancia del consumo de agua en la salud y la prevención de la enfermedad: situación actual. *Nutrición Hospitalaria*.

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*.

SJRWMD (2024). Surface water quality sampling at the st. johns river water management district. Technical report, St. Johns River Water Management District. Disponible en: https://www.sjrwmd.com/waterways/monitoring/. Accedido el 14 de mayo de 2025.

Smith, D. (1990). A better water quality indexing system for rivers and streams. *Water Resources*.

Spoto, F., Sy, O., Laberinti, P., Martimort, P., Fernandez, V., and Colin, O. (2012). Overview of sentinel-2. *2012 IEEE International Geoscience and Remote Sensing Symposium*.

Thompson, C. G., Kim, R. S., Aloe, A. M., and Becker, B. J. (2017). Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology*.

Tsai, S. M., Wang, T. N., and Ko, Y. C. (1999). Mortality for certain diseases in areas with high levels of arsenic in drinking water. *Archives of Environmental Health: An International Journal*.

USGS (2018). Secondary drinking water standards: Guidance for nuisance chemicals. Technical report, United States Geological Service. Disponible en: https://www.usgs.gov/water-science-school/science/hardness-water. Accedido el 21 de diciembre de 2025.

Vanegas, J. and Vásquez, F. (2017). Multivariate adaptative regression splines (mars), una alternativa para el análisis de series de tiempo. *Gaceta Sanitaria*.

Vanhellemont, Q. (2019). Adaptation of the dark spectrum fitting atmospheric correction for aquatic applications of the landsat and sentinel-2 archives. *Remote Sensing of Environment*.

Walfish, S. (2006). A review of statistical outlier methods. *Pharmaceutical Technology*.

Wang, Q., Shi, W., Li, Z., and Atkinson, P. (2016). Fusion of sentinel-2 images. *Remote Sensing of Environment*.

Warren, M., Simis, S., Martinez-Vicente, V., Poser, K., Bresciani, M., Alikas, K., Spyrakos, E., Giardino, C., and Ansper, A. (2019). Assessment of atmospheric correction algorithms for the sentinel-2a multispectral imager over coastal and inland waters. *Remote Sensing of Environment*.

WHO (2017). Water quality and health: Review of turbidity. Technical report, World Health Organization. Disponible en: https://www.who.int/publications/i/item/WHO-FWC-WSH-17.01. Accedido el 22 de diciembre de 2025.

WHO (2022). Arsénico. Technical report, Organización Mundial de la Salud. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/arsenic. Accedido el 14 de mayo de 2025.

Williamson, D., Parker, R. A., and Kendrick, J. (1989). The box plot: A simple visual method to interpret data. *Ann Intern Med*.