Universidad Autónoma de Querétaro

Facultad de Informática

Maestría en Ciencias de la Computación

Algoritmo estocástico con enfoque regresivo para estimar medidas del oxígeno disuelto en el agua

Tesis

Que como parte de los requisitos para obtener el Grado de

Maestro en Ciencias de la Computación

Presenta

Carlo Giovanni Cetina Camacho

Dirigido por:

M. en C. Daniel Cantón Enríquez

Co-dirigido por:

Dra. Martha Leticia Otero López

M. en C. Daniel Cantón Enríquez
Presidente

Dra. Martha Leticia Otero López
Secretario

Dr. Hugo Jiménez Hernández
Vocal

Dra. Diana Margarita Córdova Esparza
Suplente

Dr. M. Alfonso Gutiérrez López
Suplente

Centro Universitario, Querétaro, Qro.
Junio 2023
México

Algoritmo estocástico, con enfoque regresivo, para estimar medidas del oxígeno disuelto en el agua

por

Carlo Giovanni Cetina Camacho

DEDICATORIA

A mis padres, Carlos Jesús Cetina Novelo y María Angélica Camacho Ibarra, por su ayuda y consejos en los momentos buenos y malos. Me han enseñado a luchar contra las adversidades sin perder nunca la dignidad ni desfallecer en el intento. Me han dado todo lo que soy como persona.

A mis abuelos, José Pedro Camacho Solís y Graciela Ibarra Rosas, por el apoyo incondicional que me dan en la búsqueda tanto de lo personal como de lo profesional.

A mi pareja, Yessica Delgadillo Araiza, por su comprensión, cariño, y motivación durante todo el proceso de desarrollo de este trabajo de tesis.

A mi hermano Alam Yael Cetina Camacho, que es mi motivación para querer ser mejor y la persona con la que siempre puedo contar.

AGRADECIMIENTOS

A mi director de tesis, El Maestro Daniel Cantón Enríquez, por la confianza que depositó en mí. También, por su paciencia y disposición en la enseñanza y acompañamiento durante esta etapa.

A todos los profesores de la Maestría en Ciencias de la Computación, por ser guías y compartir sus conocimientos.

A mis compañeros de generación Alberto Contreras y Rafael Duarte por su compañerismo y experiencias compartidas.

A mis compañeros y compañeras del CIICCTE por su constante apoyo y orientaciones en el desarrollo de este trabajo de tesis.

A la Universidad Autónoma de Querétaro (UAQ) por el conocimiento y herramientas necesarias para seguir con mi formación profesional.

Por último, quiero agradecer al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo financiero recibido a través de la beca escolar

RESUMEN

El agua es uno de los recursos más importantes para la vida en el planeta. No obstante, la disponibilidad de esta ha ido mermando como consecuencia de la contaminación. Por lo tanto, el monitoreo de parámetros de control del agua es importante. Debido a esto, la estimación de parámetros del agua, para determinar su calidad, ha ido en aumento. Este proyecto de tesis presenta el desarrollo de un modelo que implementa un proceso estocástico basado en Regresión Lineal Múltiple (RLM) para la estimación de valores del Oxígeno Disuelto (OD). En primera instancia, se implementa la transformación de doble potencia (T2P) a los datos originales con el fin de que las variables explicativas y la variable dependiente presenten una distribución normal. Después, se realiza la separación de los datos en dos conjuntos: entrenamiento (80%) y prueba (20%). Posterior a esto, se aplicóla RLM haciendo uso del método de mínimos cuadrados con el fin de encontrar unafunción, a partir de las variables explicativas, que describa a la variable dependiente.Los parámetros de pH, Temperatura (T) y Porcentaje de Saturación del Oxígeno Disuelto (%SatOD) se usaron como variables explicativas. Por último, se implementaron las métricas de Coeficiente de Determinación (R^2) y Raíz del Error Cuadrático Medio (RECM), para evaluar el rendimiento del algoritmo propuesto.

Palabras clave: Transformada Box-Cox, Regresión Lineal Múltiple, Oxígeno Disuelto, Normalidad.

ABSTRACT

Water is one of the most important resources for life on the planet. However, its availability has been decreasing due to pollution. Therefore, the monitoring of water control parameters is important. Because of this, the estimation of water parameters to determine its quality has been increasing. This thesis project presents the development of a model that implements a stochastic process based on Multiple Linear Regression (MLR) for Dissolved Oxygen estimation. In the first instance, the double power transformation (T2P) is implemented to the original data so that the explanatory variables and the dependent variable present a normal distribution. Then, the data were separated into two sets: training (80%) and test (20%). After this, the MLR was applied using the least squares method in order to find a function, from the explanatory variables, that describes the dependent variable. The parameters pH, Temperature (T) and Dissolved Oxygen Percent Saturation (%SatOD) were used as explanatory variables. Finally, the Determination Coefficient (R2) and Root Mean Square Error (RMSE) metrics were implemented to evaluate the performance of the proposed algorithm.

Key words: Box-Cox transformation, Multiple Linear Regression, Dissolved Oxygen, Normality

Tabla de contenidos

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 INTRODUCCIÓN

El agua es uno de los recursos más importantes para la vida en el planeta, tanto para la supervivencia y desarrollo del ser humano como para los ecosistemas naturales. Sin embargo, la disponibilidad de ésta se ha ido mermando como consecuencia de la contaminación y de la necesidad de la elaboración, tanto de productos como de energía (Aquafast-FAO, 2018).

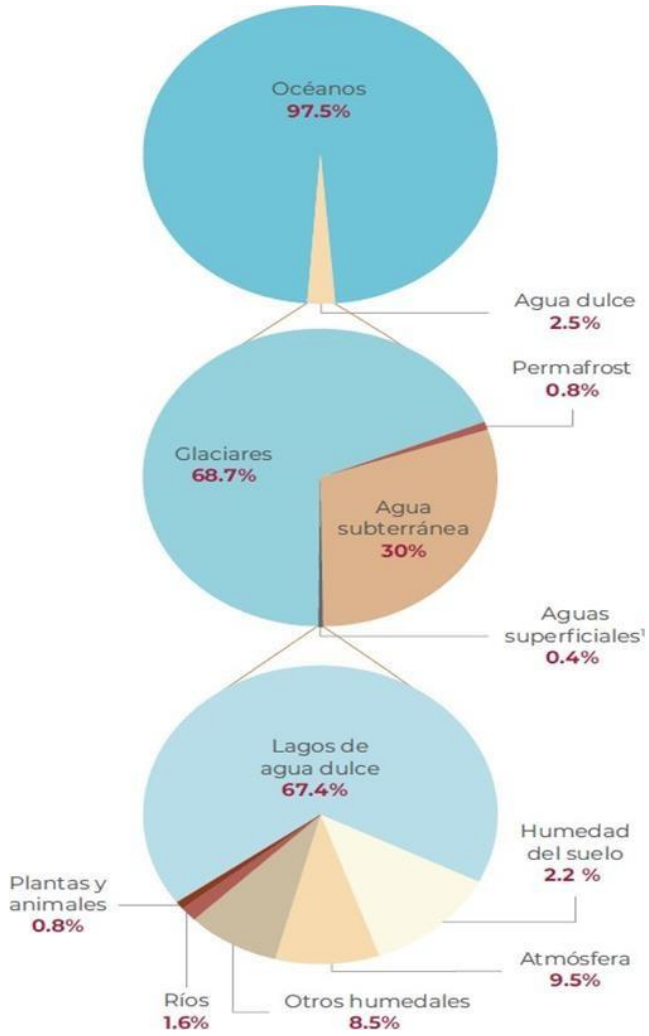
En el informe de la situación del medio ambiente en México emitido por la Secretaría de Medio Ambiente y Recursos Naturales se indica que, en todo el planeta, existen 1400 millones de kilómetros cúbicos de agua, de los cuales solo el 2.5% corresponden a agua dulce. Además, este pequeño porcentaje se localiza principalmente en ríos, lagos, glaciares, mantos de hielo y acuíferos (UNEP-GEMS, 2007). En la Figura 1, se puede observar que el mayor porcentaje de agua dulce que no está congelada lo constituye el agua subterránea con el 30%, mientras que el permafrost (suelo que tiende a formar hielo) y las aguas superficiales obtienen un 0.8 y 0.4%, respectivamente.

Por otro lado, el agua no sólo se utiliza para el consumo de seres vivos. También es importante para el funcionamiento de actividades agrícolas e industriales. Por lo que, la demanda del recurso ha ido en incremento. Este consumo consuntivo es clasificado por la Comisión Nacional del Agua (Conagua) en tres sectores principales: agrícola, abastecimiento público y sector industrial. Obteniendo un porcentaje de consumo del 70 a 90%, 7 a 18% y 1 a 11%, respectivamente, del total de consumo a nivel mundial. Debido a la alta demanda en estos sectores, la cantidad y calidad del agua se ha ido mermando y la gestión de ésta, ha obtenido un grado mayor de importancia (SEMARNAT, 2018).

Por lo anterior, es indispensable realizar un monitoreo continuo de los parámetros del agua, con el fin de obtener un reporte de la calidad del agua y tomar decisiones con base a esto.

Figura 1.

Distribución del agua en el mundo



Nota. En la figura se muestra el porcentaje del origen de agua dulce en el mundo. Fuente: SEMARNAT (2019).

1.1. Planteamiento del problema

Como posible solución, existen sistemas que realizan este monitoreo del agua mediante sensores. Conejeros-Molina et al. (2021) propone un sistema de monitoreo mediante sensores analógicos de Potencial de Hidrógeno (pH), Turbidez, Sólidos Disueltos Totales (SDT), Conductividad Eléctrica (CE) y Temperatura. Mientras que Pozo- Vásquez (2021), emplea sensores de Oxígeno Disuelto (OD), pH, CE y Temperatura.

Con las lecturas obtenidas de un sistema de monitoreo se puede determinar el grado de calidad de un cuerpo de agua por medio de índices y/o sistemas. Un

índice ampliamente utilizado es el Índice de Calidad del Agua (ICA), el cual puede determinar el grado de contaminación del agua por medio de la obtención de 18 diferentes parámetros, entre los más importantes se encuentran el Oxígeno Disuelto (OD), Demanda Bioquímica de Oxígeno (DBO_5), CE, pH y Demanda Química de Oxigeno (DQO) (SEMARNAT, 2019).

Sin embargo, la medición de parámetros resulta una tarea complicada debido a que se necesita un sensor o equipo por cada parámetro analizado. Además, llega a ser un proceso lento al necesitar personal capacitado para realizar estas mediciones. También, es importante señalar que el sensor de OD es uno de los de mayor costo, comparado con el resto de los sensores que miden los otros parámetros del agua mencionados.

En consecuencia, la inteligencia artificial, específicamente el aprendizaje automático, ha permitido en los últimos años el desarrollo de algoritmos que tienen como objetivo la estimación de ciertos parámetros del agua por medio de valores recolectados con anterioridad.

1.2. Justificación

El desarrollo de un modelo estimativo, para el monitoreo de la calidad del agua mediante el pronóstico de un parámetro por medio de otros diferentes, ha ido creciendo debido a las necesidades de los seres vivos y también, a que este tipo de modelo aprovecha datos históricos para identificar riesgos y oportunidades (Espino, 2017). Además, el uso de un enfoque de regresión para determinar la calidad del agua, tiene impactos positivos en diferentes ámbitos:

- Optimización del proceso de monitoreo de la calidad del agua (reducción de tiempo y mayor facilidad para la recolección de datos) al dejar de depender de un equipo o sensor por cada variable a medir (Carrasquilla-Batista et al., 2016).
- Reducción de costos al minimizar el número de sensores utilizados y el consumo de energía (Espino, 2017).
- Presentar datos con distribución normal para su posterior uso en el modelo

de regresión.

- Identificar la relación entre variables y brindar una herramienta para estimar la medida de una variable basándose en el conocimiento de otras (Pereira, 2010).

1.3. Objetivos

1.3.1. Objetivo General

Desarrollar un algoritmo estocástico con enfoque en regresión para estimar medidas del OD, utilizando como variables independientes los datos históricos de otros parámetros de control de contaminación del agua.

1.3.2. Objetivos Específicos

a) Transformar variables numéricas no gaussianas a variables con distribución normal para aplicar un modelo de regresión lineal.

b) Obtener una función a partir de las variables explicativas, para describir la variable dependiente.

c) Introducir los valores de prueba en la función encontrada, para estimar valores de la variable dependiente.

d) Aplicar las métricas de coeficiente de determinación y Raíz del error cuadrático medio, para evaluar el rendimiento del modelo propuesto.

e) Implementar el modelo en un prototipo de software para la construcción de librerías.

1.4. Hipótesis

Para el desarrollo de la hipótesis se planteó la siguiente pregunta de investigación: ¿Se puede generar un algoritmo estocástico con enfoque en regresión para estimar medidas del OD utilizando los datos históricos de otros parámetros del control de contaminación del agua?

Por lo que la hipótesis considerada para esta tesis es: Si se obtiene una función, de las variables explicativas, que sea capaz de describir la variable dependiente. Entonces se puede generar un algoritmo estocástico con enfoque en regresión para estimar medidas del OD.

1.5. Delimitaciones

a) Se consideran solo las muestras obtenidas en el lago Eagle Mountain en Texas, al final de verano (julio-agosto de 2019) y a 1 metro de profundidad del lago Eagle Mountain.

b) Los parámetros: Temperatura, pH y % de Saturación de Oxígeno Disuelto fueron utilizados como variables independientes. Mientras que el Oxígeno Disuelto como variable dependiente.

c) Se realiza la transformación de datos para obtener una distribución normal mediante la transformada de doble potencia.

1.6. Contribuciones

- Un artículo aceptado para ponencia en el 2do Congreso Internacional de Computación y Tecnología Educativa que se llevó a cabo del 18 al 20 de octubre de 2021 en Juriquilla, Querétaro, México. El título del artículo es: “Métodos predictivos para el monitoreo de oxígeno disuelto en ecosistemas acuáticos”.

- Un certificado de registro público de derecho de autor en la rama de programas de cómputo con número de registro: 03-2023-091210350400-01, titulado “Software para la estimación de medidas del oxígeno disuelto mediante un proceso estocástico”.

- Participación en el 1er, 2do y 3er Coloquio Nacional en Computación y Aplicaciones Tecnológicas, con la ponencia “Algoritmo estocástico con enfoque regresivo para estimar medidas del oxígeno disuelto en el agua”.

1.7. Estructura de tesis

La estructura del presente trabajo de tesis está organizada de la siguiente manera: En el Capítulo 2, se presentan los antecedentes y fundamentación teórica, encontrados en el estado del arte, acerca de los modelos, algoritmos y métodos de regresión utilizados para la estimación de parámetros de control del agua. En el Capítulo 3, se describe la metodología utilizada para el desarrollo del algoritmo: el área de estudio y recolección de los datos, transformación de datos, desarrollo del algoritmo estocástico y su validación. En el Capítulo 4, se describen los resultados obtenidos previos y posteriores a la transformación de datos, así como, una comparativa de los trabajos consultados y este trabajo. Por último, en el Capítulo 5 se mencionan las conclusiones y las sugerencias de trabajos a futuro.

2. ANTECEDENTES

Los modelos con enfoques de regresión suelen tratar problemas de predicción cuando se trata de variables cuantitativas continuas, como es el caso del OD.

Carrasquilla-Batista (2016) hace uso de la Regresión Lineal Simple (RLS), la cual se centra en explicar la relación que tiene una variable independiente o predictora con la variable dependiente por medio de una ecuación que representa la línea que mejor se ajusta a los puntos obtenidos de las variables. Además, en eltrabajo mencionado se desarrolló un modelo basado en este método para lapredicción de parámetros del agua en un entorno de producción de microalgas. También, se usaron como variables independientes (por separado) el pH, temperatura y OD.

Existen casos, como los que se presentan a continuación, en los que se hace la suposición de que existe más de un factor que puede afectar a la variable dependiente, por lo que se opta por desarrollar un modelo basado en Regresión Lineal Múltiple (RLM). Ewaid (2018) utilizó inicialmente 23 parámetros hidrológicos como variables independientes. Entre los que destacan el OD, pH, Conductividad

Eléctrica, Temperatura, DBO_5, Turbidez y SDT. Por otra parte, se trata como variable dependiente al ICA, con el fin de obtener un pronóstico y saber si el agua del río Bagdad, en Iraq, era apta para consumo humano.

Los métodos sobre la predicción de la calidad del agua no sólo son empleados en determinar si el líquido es apto para consumo humano, sino también son aplicados en entornos industriales (Abyaneh, 2014) y de agricultura (Yang, et al., 2017). El primero propone dos modelos usando RLM con el fin de aplicarlos para la predicción de DBO_5 y DQO en aguas residuales generadas por la industria. Además, usa como variables regresoras a la temperatura, pH y el SDT.

A su vez, Yang et al. (2017) usaron el mismo método para realizar pronósticos en estanques de producción acuícola. No obstante, la variable dependiente es distinta y la calidad del agua se midió con base a la cantidad de OD presente. De ahí que, se agregan variables independientes como la CE, salinidad y OD y otras no son tomadas en cuenta, como los SDT.

Es necesario señalar que hay veces en las que la relación entre variables está lejos de ser lineal y esto puede presentar restricciones en la capacidad de predecir del modelo. Por lo que, en estos casos, se puede llegar a elegir un método de regresión no lineal con el fin de obtener un mejor ajuste.

Entre estos métodos se encuentra la Regresión Polinomial (RP), la cual tiene mejor adaptabilidad debido a que se le agregan potencias, por lo que la función que representa el comportamiento de los datos puede llegar a ser un polinomio cuadrado, cúbico, de grado 4, etc.

En el trabajo de Ahmed (2019) se presentó el uso de RP de grado 2 para la predicción del ICA por medio de parámetros de entrada como la temperatura, pH, turbidez y el SDT. También se resaltó que se debe de tener cuidado al momento de escoger el grado del polinomio, debido a que el buscar tener un error menor

elevando el grado del polinomio, puede resultar en un sobre ajuste.

Un hecho que deben cumplir los datos para la aplicación de regresión lineal es que deben de tener un comportamiento normal en su distribución. Sin embargo, dicha distribución no se consigue en la mayoría de los casos. En todos los trabajos consultados los autores suponen que los datos utilizados tienen una distribución normal. No obstante, si no se cumple lo anterior, los resultados de estas pruebas se vuelven poco confiables y no se pueden generalizar los hallazgos.

Por otro lado, previo a la aplicación de un método de regresión, se puede realizar una transformación de los datos. En Gupta (1987), se utilizó la transformada de doble potencia para que la distribución de los datos tienda a una normal, cuando el coeficiente de sesgo se aproxime a 0 y el coeficiente de curtosis se aproxime a 3.

Riani et al. (2022) hace énfasis en la determinación automática de transformaciones paramétricas en modelos de regresión para la aproximación a la normalidad. Considerando la transformación Box- Cox y su generalización a la transformación extendida de Yeo-Johnson, esto debido a que permite respuestas tanto positivas como negativas.

Así mismo, Nwakuya y Anyaogu (2022) presentan el uso de la transformación Yeo-Johnson para la aplicación de una regresión cuantílica sin y con transformación. Los resultados demuestran que el modelo con transformación tiene un mejor desempeño que el modelo al que no se le aplicó transformación.

El trabajo de Efremides et al. (1994) es el único encontrado en el estado del arte en el que se aplica una transformación de los datos para la aplicación de un modelo estocástico. Dicho trabajo presenta una propuesta para la simulación de la secuencia de precipitaciones utilizando el método de matrices de transición. Sin embargo, las muestras mostraron que la mayoría de los valores estaban muy sesgados y tendían a agruparse hacia el límite inferior, por lo que se optó por aplicar la transformada de doble potencia para eliminar dicho sesgo y obtener una distribución normal.

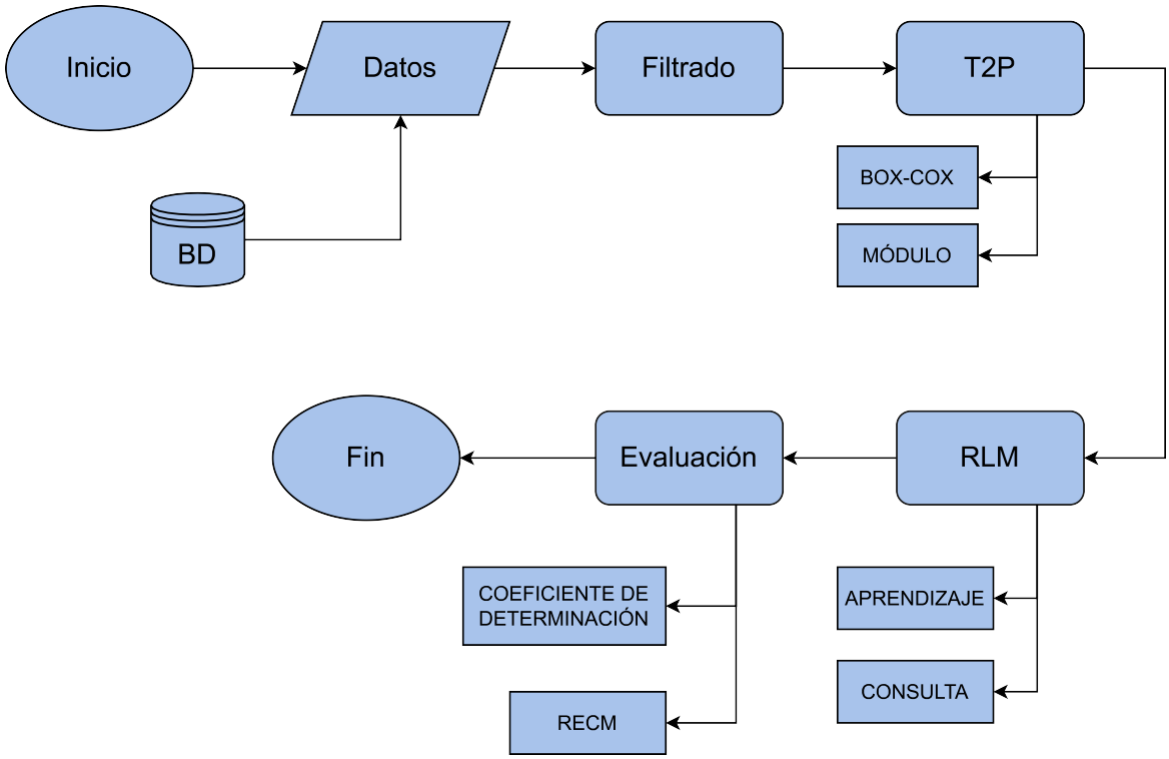
En la revisión de la literatura se encontró un único trabajo que presenta una transformación de los datos antes de la aplicación del modelo. No obstante, el presente trabajo de tesis considera la transformación de los datos para el uso del modelo de Regresión Lineal Múltiple con el fin de obtener medidas del Oxígeno Disuelto, por medio de otras variables de control de contaminación del agua, en un entorno acuático.

3. METODOLOGÍA

En este capítulo se describe el procedimiento utilizado para el cumplimiento de los objetivos planteados en el presente trabajo de investigación. Así mismo, se presenta la descripción del lugar de estudio y de los parámetros utilizados en el algoritmo de regresión. La Fig. 2 muestra el orden de las fases planteadas acompañadas de una descripción.

Figura 2

Esquema de la metodología propuesta



Fuente: Elaboración propia.

3.1. Recopilación y filtrado de datos

El conjunto de datos usado en esta tesis fue tomado a partir del trabajo presentado por Durell et al. (2022). Este conjunto de datos fue recolectado por la institución Tarrant Regional Water District y las muestras se obtuvieron del lago Eagle Mountain en Texas. Los datos se recolectaron cada 2 horas en 21 diferentes profundidades, medidas cada 0.5 metros empezando desde la superficie y terminando en los 10 metros de profundidad. El conjunto de datos está conformado por un total de 2252 mediciones de 4 diferentes parámetros de control de contaminación del agua, tomadas entre el 25 de abril del 2019 y el 29 de octubre del mismo año.

Los parámetros medidos fueron: la temperatura (*T*), que es medida en grados Celsius; el pH (*pH*), que indica la alcalinidad/acidez del agua, va desde 0 (ácida) hasta 14 (alcalina). El porcentaje de Saturación de Oxígeno Disuelto (*%SatOD*), que es el porcentaje de *OD* relativo a la concentración en equilibrio con la atmósfera. *%SatOD* puede variar de 0% a 200%, donde *%SatOD* de 100% indica que el porcentaje de *OD* en la atmósfera es el mismo que el porcentaje de OD en el agua. Los valores de *%SatOD* por encima del 100% son causados por la turbulencia del viento sobre el lago y la fotosíntesis de las plantas acuáticas. Los valores por debajo del 100% son causados por la respiración de animales acuáticos y descomposición en el fondo del lago.

Las mediciones que conforman el conjunto de datos se separaron según la época del año, teniendo como resultado 3 bloques diferentes. El primero está conformado por las mediciones recolectadas en los meses de abril, mayo y junio (principios de verano). El segundo, en los meses de julio y agosto (finales de verano). El último, en los meses de septiembre y octubre (principios de otoño). No obstante, en el presente trabajo se optó por hacer uso de las muestras recolectadas a finales de verano a 1 metro de profundidad. Esto debido a que en los bloques restantes existe más concentración en el agua a causa de las sequías.

Por lo que se usaron un total de 800 muestras de las cuales el 80% fueron destinadas a la etapa entrenamiento y 20% a la etapa de validación.

3.2. Descripción del lugar de estudio

El lago Eagle Mountain se encuentra en el norte de Texas. En la Fig. 3 se muestra un mapa geográfico del lago y sus alrededores. El lago abarca un aproximado de 35Km^2 y es usado para fines recreativos, pesca y como suministro de agua.

Figura 3

Mapa geográfico del lago en la montaña Eagle, Texas

Fuente: Tarrant Regional Water District. https://www.trwd.com/eagle-mountain-lake/

3.3. Transformación de los datos

El uso de RLM presenta ciertas características o supuestos previos a su aplicación. Wapole (2012) menciona que uno de estos supuestos es que la distribución observada de las variables, tanto explicativas como dependientes, debe de ser de tipo normal. Para satisfacer dicho supuesto para un análisis paramétrico como la regresión lineal, el coeficiente de asimetría se debe reducir a un valor cercano a 0 y el coeficiente de curtosis debe alcanzar un valor cercano al 3 (Gupta etal., 1987).

3.3.1. Coeficiente de asimetría (sesgo)

El sesgo mide el grado de asimetría de una distribución con respecto a su media. Si $As > 0$, entonces la distribución se encuentra sesgada a la izquierda. Si $As < 0$, la distribución está sesgada a la derecha.

Se determinó el sesgo de la distribución de cada una de las variables, haciendo uso de la fórmula del coeficiente de asimetría de Fisher, la cual está definida por la Ecuación 3.1.

$$As = \frac{\sum_{i=1}^{N}(x_i-\mu)^3}{N \cdot \sigma^3} \tag{3.1}$$

3.3.2. Curtosis

El coeficiente de curtosis analiza el grado de concentración que presentan los valores, de un conjunto de datos, alrededor de la zona central de la distribución. Por lo que, un valor alto en el coeficiente indica una distribución apuntada, mientras que un valor bajo indica una distribución achatada. Existen tres tipos de curtosis:

- Leptocúrtica: Es una distribución muy apuntada. Obtiene un valor muy alto en el coeficiente de curtosis.
- Mesocúrtica: El coeficiente de curtosis es equivalente al de la distribución normal, obteniendo un valor cercano a 3. No se considera ni apuntada ni achatada.
- Platicúrtica: Es una distribución muy achatada. La concentración alrededor de la zona central es baja.

El coeficiente de curtosis de cada variable se determinó mediante la fórmula

que está definida por la Ecuación 3.2.

$$C = \frac{\sum_{i=1}^{N}(x_i-\mu)^4}{N \cdot \sigma^4} \quad (3.2)$$

3.3.3. Transformación de doble potencia (T2P)

En el uso análisis de paramétricos, como lo es la regresión lineal, se suelen hacer suposiciones tentativas que no son ciertas pero que se cree que pueden ser

útiles. Por ejemplo, la suposición de que la distribución de los datos tiene un comportamiento normal. Sin embargo, el no corroborar este hecho y en caso de no cumplirse, puede llevar a pruebas que se vuelven poco confiables y a obtener hallazgos que no se pueden generalizar.

Debido a esto, se propuso realizar la transformación de los datos originales para obtener una distribución normal y posterior a esto, realizar el análisis de RLM.

En primera instancia, se aplicó la Transformada de Box-Cox, La cuál se describe en la Ecuación 3.3. Box-Cox se usa para transformar variables no normales a una forma normal.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(\lambda), & \lambda = 0 \end{cases} \quad (3.3)$$

Para la definición del valor de lambda, se encontró el valor óptimo mediante un análisis de fuerza bruta en un intervalo dado. El valor óptimo de lambda, y su sustitución dentro de la función de Box-Cox, permite obtener un coeficiente de asimetría cercano a 0, por lo que el histograma de las variables es simétrico.

Sin embargo, a pesar de que el coeficiente de asimetría estaba cercano al valor necesario, los valores de curtosis estaban lejos de ser 3. Y, como se había comentado con anterioridad, la distribución sería verdaderamente normal con valores de los coeficientes cercanos a $As = 0$ y $C = 3$.

Cuando la curtosis no es un valor cercano a 0, el histograma puede ser apuntado (Leptocúrtica) o plano (Platicúrtica). El valor de la curtosis fue corregido mediante la transformación módulo, la cual está definida por la Ecuación 3.4.

$$t_i = |y_i - \mu|^{\gamma} \quad (3.4)$$

Donde μ es el promedio de la serie transformada con Box-Cox; y γ es el parámetro de la transformación módulo, donde el valor óptimo de γ permite obtener un coeficiente de curtosis cercano a 3.

3.4. Regresión Lineal Múltiple

La regresión lineal es un método de modelado estadístico usado para la explicación de la relación entra una o más variables independientes y una variable dependiente. No obstante, existen problemas de investigación en los que se necesita más de una variable independiente para el modelo. Cuando un modelo es lineal en sus coeficientes y tiene más de una variable independiente, se le denomina modelo de Regresión Lineal Múltiple (RLM) y la respuesta estimada se obtiene a partir de la Ecuación 3.5.

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k \tag{3.5}$$

En donde cada coeficiente de regresión b_k se estima a partir de los datos muestrales, usando el método de los mínimos cuadrados.

3.4.1. Enfoque matricial del modelo de regresión lineal

El conocimiento sobre la teoría de matrices facilita la resolución en las ecuaciones que describen los valores de respuesta en el modelo de RLM (Ronald et al. 2012). Al usar la notación de matrices, se puede obtener la Ecuación 3.6

$$y = X\beta \tag{3.6}$$

donde:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_3 \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Si se usa el concepto de mínimos cuadrados para obtener los estimados de β, la expresión se puede minimizar y se obtiene la Ecuación 3.7.

$$SCE = (y - Xb)'(y - Xb) \tag{3.7}$$

Este proceso de minimización implica resolver b para la Ecuación 3.8.

$$\frac{\partial}{\partial b}(SCE) = 0 \tag{3.8}$$

El resultado se reduce a la solución de b en la Ecuación 3.9.

$$(X'X)b = X'y \tag{3.9}$$

Si la matriz $(X'X)$ es no singular, la solución para los coeficientes de regresión se escribe con le Ecuación 3.10.

$$b = (X'X)^{-1}X'y \tag{3.10}$$

3.5. Métricas de evaluación

Las métricas utilizadas para la etapa de evaluación del algoritmo son el coeficiente de determinación (R^2) y la Raíz del Error Cuadrático Medio (RECM).

El primero se define como la porción de variabilidad de la variable dependiente que es explicada por la regresión, se expresa con la Ecuación 3.11.

$$R^2 = \frac{\sum_1^N (\hat{Y} - \bar{y})^2}{\sum_1^N (Y - \bar{y})^2} \tag{3.11}$$

En cuanto al RECM, permite cuantificar la cercanía entre los valores observados y los valores estimados del modelo, se denota con la Ecuación 3.12.

$$RECM = \sqrt{\frac{\sum_{1}^{N}(\hat{Y}-Y)^2}{N}} \quad (3.12)$$

4. RESULTADOS

El algoritmo desarrollado en la actual investigación se implementó en Matlab R2020a

En este capítulo se exponen los resultados obtenidos a partir de la implementación del algoritmo a un conjunto de datos de parámetros de control de contaminación del agua. En primera instancia, se presentan los histogramas y los valores de los coeficientes de asimetría y curtosis, tanto de los datos sin transformar como de losdatos resultantes de la aplicación de la transformada de doble potencia. Posterior a esto, se describe el nivel de bondad que tiene el ajuste del modelo de Regresión Lineal Múltiple mediante los valores del Coeficiente de Determinación y el RECM. Después, se analizan la distribución presentada por los residuos resultantes del análisis de regresión. Y por último, se realiza una discusión y comparación de los trabajos consultados.

4.1. Análisis previo de los parámetros de control del agua

En la Fig. 4**,** se muestra la distribución de frecuencias de los valores originales de las variables temperatura, oxígeno disuelto, porcentaje de saturación de oxígeno disuelto y pH. De forma cualitativa, se observa que las distribuciones presentan una asimetría con un sesgo a la izquierda. Lo que significa que se tiene

una curva de asimetría positiva y que muchos de los datos son menores que la media. En cuanto a la curtosis, se puede observar que la temperatura presenta una distribución relativamente elevada y el pH de una distribución relativamente plana.

En la Tabla 1, se indica de forma cuantitativa el valor de asimetría que presenta cada una de las variables. Por lo que se puede inferir que, efectivamente, la distribución que se presenta; en cada uno de los histogramas, tiene un ligero sesgo hacia la izquierda. Siendo el %SatOD la variable que presenta el coeficiente de asimetría más elevado y el pH la variable con el valor más cercano al 0. Así mismo, se puede corroborar que la distribución de la T es la que presenta un valor mayor en el coeficiente de curtosis; siendo la única que tiene un dicho coeficiente un valor por encima de 3. Por otra parte, la distribución del pH, OD y %SatOD se presenta platicúrtica (valor por debajo de 3), lo que muestra que hay una menor concentración de datos en torno a la media.

Figura 4

Histogramas de los parámetros físicos y químicos de control del agua estudiados en el lago Eagle Mountain, Texas.

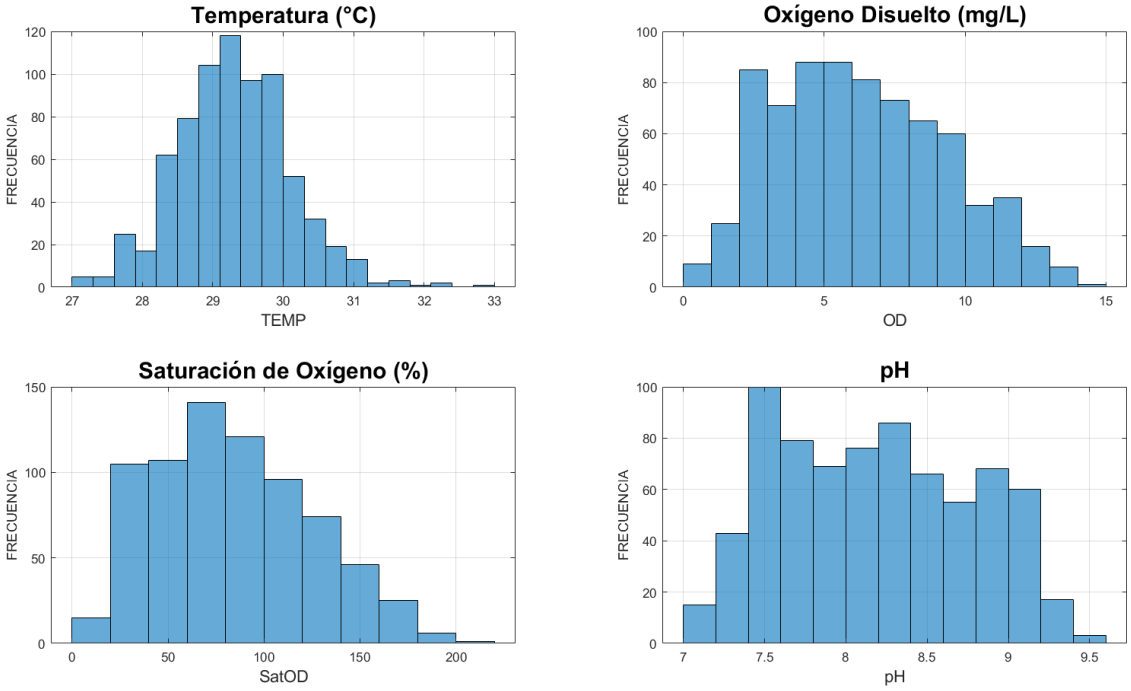


Tabla 1

Coeficiente de asimetría y de curtosis de los valores originales de cada variable.

Variable	Coef. Asimetría	Coef. Curtosis
T	0.2768	3.7086
OD	0.3195	2.3500
%SatOD	0.3824	2.4308
pH	0.1634	1.9169

4.2. Transformada de doble potencia

Uno de los supuestos que se debe cumplir, para la aplicación de RLM, es que los datos presenten una distribución normal. El comportamiento normal en una distribución exige que el coeficiente de curtosis sea igual 3 (distribución mesocúrtica) y el coeficiente de asimetría sea 0.

Los coeficientes de las variables utilizadas presentan valores diferentes a los necesarios. Por lo que, antes de realizar RLM se aplicó la transformada de doble potencia a las diferentes variables. En las Fig. 5 y 6 se presentan las distribuciones de las variables en sus diferentes etapas en la transformación de doble potencia. Primero, se muestra el histograma con los datos originales de cada variable. Después, se observan los datos transformados haciendo uso, únicamente, de Box-Cox. Por último, el resultado de la aplicación de la transformada de doble potencia (Box-Cox / Módulo).

Se observa que, con la primera transformación, el sesgo hacia la izquierda, que presentaban los datos originales, se reduce considerablemente. Sin embargo, la concentración de la mayoría de los datos sigue sin estar alrededor de la media. Tal es el caso del pH, que sigue mostrando una distribución uniforme. A diferencia de la transformación utilizando sólo Box-Cox, la transformación de doble potencia

presenta mejorías tanto en el sesgo como en la curtosis.

En la Tabla 2 se muestran los valores obtenidos en ambos casos. Se puede verificar que la transformada de Box-Cox obtiene mejores resultados en cuanto al sesgo. Sin embargo, la transformada de doble potencia presenta una mejoría, en contraste con los datos originales, tanto en sesgo como en curtosis.

Figura 5

Histogramas de los datos originales y transformados de las variables de T y OD.

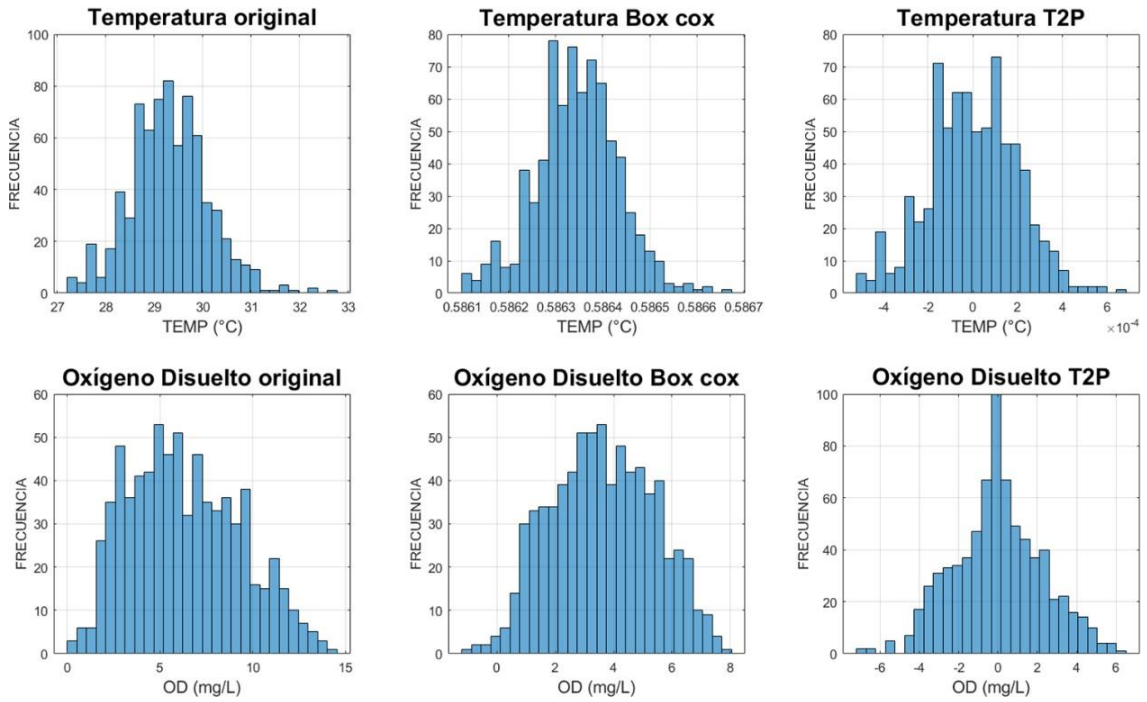


Figura 6

Histogramas de los datos originales y transformados de las variables de %SatOD y pH.

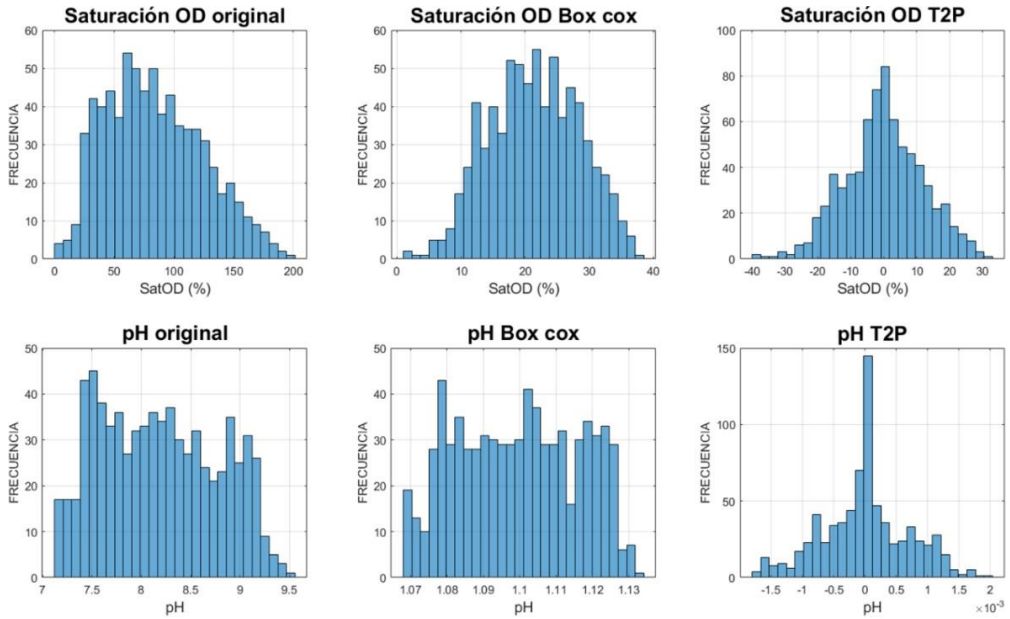


Tabla 2

Coeficiente de asimetría y curtosis de los datos transformados utilizando Box-Cox y T2P.

	Box – Cox		T2P (Box-Cox/Módulo)	
Variable	**Coef. Asimetría**	**Coef. Curtosis**	**Coef. Asimetría**	**Coef. Curtosis**
T	-0.0038	3.4233	**-0.0132**	**2.9902**
OD	0.0210	2.3153	**0.0138**	**3.0116**
%SatOD	-0.0391	2.3929	**-0.0760**	**3.0034**
pH	0.0003	1.8889	**-0.0259**	**3.0047**

4.2. Análisis de regresión lineal múltiple con datos transformados

En la Tabla 3 se muestra el análisis de regresión lineal múltiple, con los datos transformados (RLM-T2P) para la estimación del OD en el lago Eagle Mountain, haciendo uso de un conjunto de 600 muestras para el entrenamiento. El R^2 alcanza

un valor de 0.9992 y un RECM de 0.0582. La bondad del ajuste sugiere que las variables independientes tienen un porcentaje elevado de estimación sobre la variable dependiente. Así mismo, con el valor de RECM se determina que la distancia ente los valores observados y los valores estimados es corta.

Tabla 3.

Regresión lineal múltiple usando los datos transformados.

Modelo	$\mathbf{R^2}$	RECM
RLM-T2P	0.9992	0.0582

El resultado sugiere que el modelo de regresión lineal múltiple con los datos transformados, logra una elevada precisión. En la Fig. 7 y la Tabla 4 se observa el comportamiento de la distribución de los residuos, así como el coeficiente de asimetría y de curtosis. Los residuos se definen como la diferencia entre los datos reales y los datos calculados del modelo. Después de obtener los coeficientes, se puede inferir que los residuos presentan una distribución normal.

Figura 7

Histograma de los residuos del modelo RLM-T2P.

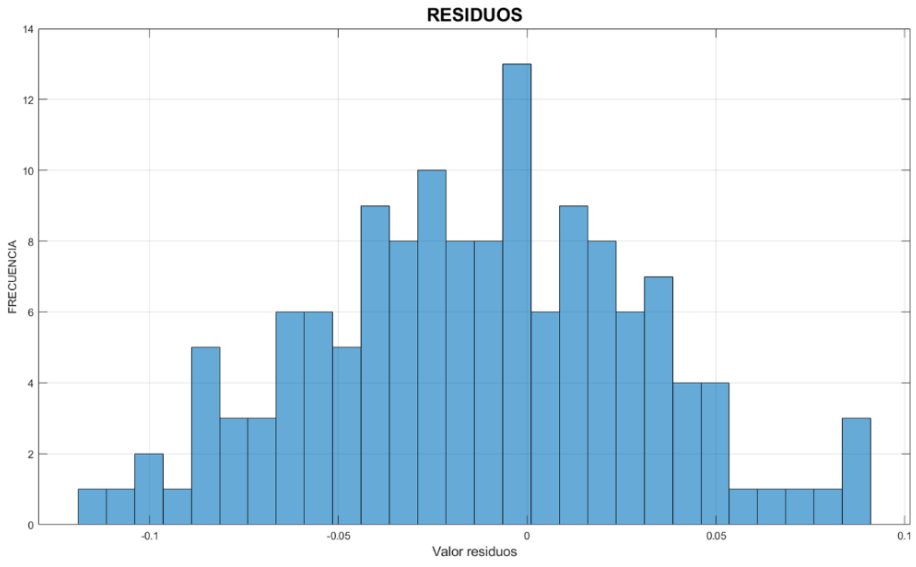


Tabla 4.

Coeficiente de asimetría y de curtosis de los residuos del modelo RLM-T2P

	Coef. Asimetría	**Coef. Curtosis**
Residuos	0.0097	2.8602

4.3. Comparación con los trabajos consultados

En la Tabla 5 se exponen trabajos consultados utilizados como referencia y/o comparación en el desarrollo de esta propuesta de tesis. Los autores de dichos trabajos buscan predecir el valor de un parámetro con el fin de determinar la calidad del agua por medio de la regresión. Por otro lado, los trabajos consultados se clasifican conforme a seis características.

Tabla 5

Trabajos consultados con respecto al algoritmo estocástico propuesto.

Trabajo	**Método**	**No. Parámetros entrada**	**Parámetros salida**	**Distribución**	**Métrica**	**Transformación de Datos**
(Seeboonruang ,2017)	RLM	5	CE	Se asumen gaussianos	R^2 (0.899)	N/A
(Yang, et al., 2017)	RLM	5	OD	Se asumen gaussianos	*r (0.9423)*	N/A
(Ewaid, et al., 2018)	RLM	5	ICA	Se asumen gaussianos	R^2 (0.974)	N/A.
(Ahmed, et al., 2019)	RP	4	ICA	N/A	R^2 (0.6573) RECM (3.4286)	N/A
(Yildiz, et al.,2019)	RLM	4	ICA	Se asumen gaussianos	R^2 (0.6)	N/A

(El Bilali, et al.,2020).	RLM	2	ICA	Se asumen gaussianos	r (0.96)	N/A
(Gil Marín, 2020)	N/A	5	ICA	Se asumen gaussianos	R^2 (0.99)	N/A
(Gaya, et al., 2020)	RLM	6	ICA	Se asumen gaussianos	R^2 (0.8919) RECM (0.0362)	N/A
(Hsu, et al.,2020)	RP	3	OD	N/A	SCE (0.82%)	N/A
(Mundi, 2021)	RLM	3	ICA	N/A	RECM (11.25)	N/A
Propuesta actual	RLM	3	OD	No gaussianos	R^2 (0.9992) RECM (0.0582)	T2P

Fuente: elaboración propia.

Se concluye de acuerdo con el análisis de los trabajos consultados lo siguiente:

a) Existen diversos métodos que tienen como enfoque la regresión. En la Tabla 5 se observa que el más empleado para la predicción de variables en cuanto a la calidad del agua es la RLM. No obstante, se debe señalar que también es utilizada la RP.

b) La cantidad parámetros de entrada varía entre los diferentes trabajos. Pero al ser RLM el método más utilizado, las variables explicativas tienden a ser 3 o más.

c) En cuanto a los parámetros de salida, más de dos tercios de los trabajos realizan la estimación del Índice de Calidad del Agua. Mientras que el Oxígeno Disuelto pretende ser estimado por el tercio restante de los trabajos consultados.

d) Los trabajos que aplicaron RLM como método de estimación asumen que existe una distribución normal en los datos utilizados. Por su parte, los trabajos de Regresión Polinomial no mencionan dicho supuesto debido a que no es requisito para este método. El presente trabajo es el único que corrobora la no normalidad de los datos.

e) Las métricas más usadas para trabajos de estimación son el coeficiente de determinación y la Raíz del Error Cuadrático Medio. En el que, para el primer coeficiente, el presente trabajo presenta el valor más alto con 0.992. Seguido del trabajo realizado por Gil Marín (2020), con un valor de 0.990 Por otro lado, el trabajo de Gaya et al. (2020) presenta el error más bajo en la estimación de sus valores (0.0362), seguido del actual trabajo (0.582).

f) Todos los trabajos hacen la suposición de que los datos que se utilizan tienen una distribución normal. Por lo que no presentan un algoritmo para la transformación de los datos, con excepción del presente trabajo. En el cual, se utilizó la transformada de doble potencia.

5. CONCLUSIONES Y TRABAJOS A FUTURO

5.1. Conclusiones

El presente trabajo de investigación se expone un modelo que permite la estimación del Oxígeno Disuelto por medio de otras variables de control de contaminación del agua obtenidas del lago Eagle Mountain, Texas. Esta propuesta emplea el bajocoste computacional del método de Regresión Lineal Múltiple, así como su capacidad para trabajar con más de una variable dependiente para explicar la variable independiente.

Con base en los resultados obtenidos, se observa una mejora de forma cualitativa y cuantitativa en los valores en los coeficientes de asimetría y curtosis. Por ejemplo, en el caso del pH, la distribución original era uniforme y con un pequeño sesgo a la izquierda. No obstante, se aplicó la transformada de Box-Cox y se eliminó dicho sesgo. Sin embargo, la distribución siguió teniendo un comportamiento uniforme, y no cambió hasta que se le aplicó la transformada módulo. Esto permite cumplir con la presunción de la normalidad de los datos y poder someter las variables transformadas a un análisis de Regresión Lineal Múltiple.

Por otro lado, el presente trabajo de tesis cumple con los objetivos establecidos, ya que se desarrolló un algoritmo estocástico capaz de transformar variables no gaussianas a variables con distribución normal con el fin de obtener una función que explica la variable dependiente por medio de la Regresión Lineal Múltiple y obtener estimaciones de valores del Oxígeno Disuelto haciendo uso de otras variables de control de contaminación del agua.

De acuerdo con la revisión del estado del arte, el presente trabajo es el único que cumple con la presunción de la normalidad de los datos para la aplicación de RLM. Lo anterior, se cumple debido a la transformación de doble potencia.

Con la hipótesis presentada en este trabajo de investigación se logró comprobar a que la función encontrada a partir del pH, la Temperatura y el % de Saturación de Oxígeno Disuelto fue capaz de explicar el comportamiento y con ello , realizar estimaciones de Oxígeno Disuelto. También, es necesario mencionar que el algoritmo desarrollado en este trabajo puede aplicarse a otras áreas de estudio.

5.2. Trabajos a futuro

A continuación, se describen algunos de los trabajos a futuro por desarrollar:

- Realizar un análisis de la estacionalidad en los datos con la finalidad de utilizar el algoritmo desarrollado de forma no estacional en periodos de estacionariedad.
- Incluir un análisis multivariable para trabajar con el número óptimo de variables independientes.
- Optimización de parámetros del algoritmo, con el fin de escoger el parámetro óptimo de forma automática y no por fuerza bruta.

6. REFERENCIAS

Aquastat-FAO. Sistema de información sobre el uso del agua en la agricultura y el medio rural de la FAO. Disponible en: http://www.fao.org/nr/water/aquastat/main/index.stm. Fecha de consulta: agosto de 2021.

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. Water, 11(11), 2210. doi:10.3390/w11112210

Ambiental y de Crecimiento Verde. Ciudad de México (México): Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT), Edición 2018, Cap. 6, Agua (pp. 379-449).

Avila, R., Horn, B., Moriarty, E., Hodson, R., & Moltchanova, E. (2018). Evaluating statistical model performance in water quality prediction. Journal of Environmental Management, 206, 910–919.

Carrasquilla-Batista, Arys, Chacón-Rodríguez, Alfonso, Núñez-Montero, Kattia, Gómez- Espinoza, Olman, Valverde, Johnny, & Guerrero-Barrantes, Maritza. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. Revista Tecnología en Marcha, 29(Suppl. 5), 33-45.

Conejeros Molina, Alvaro, Hueichaqueo Pichunman, Camilo, Martinez-Jimenez, Boris L., & PlaceresRemior, Arley. (2021). Monitoreo de calidad del agua en sistema

de agua potable rural. Ingeniería Electrónica, Automática y Comunicaciones, 42(3), 60-70. Epub 11 de diciembre de 2021.

Durell, L., Scott, J. T., Nychka, D., & Hering, A. S. (2022). Functional forecasting of dissolved oxygen in high-frequency vertical lake profiles. Environmetrics, e2765.

Efremides, D., Tsakiris, G (1994).. Stochastic modelling of point rainfall in a Mediterranean island environment. Water Resour Manage 8,. 171–182 https://doi.org/10.1007/BF00877085

El Bilali, A., & Taleb, A. (2020). Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment. Journal of the Saudi Society of Agricultural Sciences.

Espino Carlos. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso. [Grado en ingeniería informática, Universitat Oberta de Catalunya]. Repositorio InstitucionalUOC.http://openaccess.uoc.edu/webapps/o2/bitstream/10609/59565/6/ caresptimTFG0117mem%C3%B2ria.pdf

Ewaid, S. H., Abed, S. A., & Kadhum, S. A. (2018). Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis. Environmental Technology & Innovation, 11, 390–398.

Gaya, M.S., Abba, S.I., Abdu, A.M., Tukur, A.I., Saleh, M.A., Esmaili, P., & Wahab, N.A. (2020). Estimation of water quality index using artificial intelligence approaches and multi-linear regression. IAES International Journal of Artificial Intelligence, 9, 126- 134.

Gil Marin, J. A. (2020). Modelo de calidad del agua subterránea mediante el uso combinado del análisis de componentes principales (ACP) y regresiones lineales múltiples (RLM): Caso de estudio: acuíferos de Maturín, Monagas, Venezuela. INNOTEC, (20 jul-dic), 67–88.

Gómez, I. & Peñuela, G. (2016). Revisión de los métodos estadísticos multivariados usados en el análisis de calidad de aguas. Revista Mutis6(1), 54-63.

González Vidal Ana. (2015). Selección de variables: Una revisión de métodos existentes. 2021, de Universidad de La Coruña Sitio web: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1263.pdf

Gupta, D.K., Asthana, B.N., Bhargawa, A.N. (1987). Estimation of Design Flood. In: Singh, V.P. (eds) Application of Frequency and Risk in Water Resources. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-3955-4_8

Hsu, W.-C., Chao, P.-Y., Wang, C.-S., Hsieh, J.-C., & Huang, W. (2020). Application of Regression Analysis to Achieve a Smart Monitoring System for Aquaculture. Information, 11(8), 387.

Huang, H., Wang, Z., Xia, F. et al. Water quality trend and change-point analyses using integration of locally weighted polynomial regression and segmented regression. Environ Sci Pollut Res 24, 15827–15837 (2017).

Kadam, A. K., Wagh, V. M., Muley, A. A., Umrikar, B. N., & Sankhua, R. N. (2019). Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. Modeling Earth Systems and Environment, 5(3), 951–962.

Mundi, G., Zytner, R. G., Warriner, K., Bonakdari, H., & Gharabaghi, B. (2021). Machine Learning Models for Predicting Water Quality of Treated Fruit and Vegetable Wastewater. Water, 13(18), 2485. MDPI AG.

Nwakuya, M. T., & Anyaogu, I. V.(2022) Implementation of Yeo-Johnson Transformation in Quantile Regression.

Pereira Augusto. (2010). ANALISIS PREDICTIVO DE DATOS MEDIANTE TECNICAS DE REGRESION ESTADISTICA. [Máster en investigación informática, Universidad Complutense de Madrid]. Repositorio Institucional UCM.
https://eprints.ucm.es/id/eprint/11389/1/Analisis_Predictivo_de_Datos.pdf

Ronald E. Walpole, Raymond H. Myers, Sharon l. Myers y Keying Ye. (2012). Regresión lineal simple y correlación, Probabilidad y estadística para ingeniería y ciencias, (Novena ed., pp. 389-450) Pearson.

Pozo Vásquez, J. D. (2021). Prototipo para monitoreo de la calidad de agua de riego. Guayaquil [Tesis de Licenciatura, Gayaquil] Repositorio Institucional de la Universidad de Guayaquil http://repositorio.ug.edu.ec/handle/redug/51623

Riani, M., Atkinson, A.C. & Corbellini, A. Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression. Stat Methods Appl 32, 75–102 (2023).

Seeboonruang, U. (2017). A Multiple Regression Analysis for Predicting Salinity in Shallow Groundwater. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 2, pp. 15-17).

SEMARNAT (2018). Informe de la Situación Del Medio Ambiente en México. Compendio de Estadísticas Ambientales. Indicadores Clave, de Desempeño.

SEMARNAT (2019). Indicadores de calidad del agua. Gerencia de calidad del agua. http://dgeiawf.semarnat.gob.mx:8080/ibi_apps/WFServlet?IBIF_ex=D3_R_AGUA05_01%26IBIC_user=dgeia_mce%26IBIC_pass=dgeia_mce#:~:text=Indicadores%20de%20calidad%20del%20agua&text=Este%20sistema%20se%20denomin%C3%B3%20%C3%8Dndice,un%20porcentaje%20de%20agua%20pura. 35(12), 2881–2894. doi:10.1016/s0043-1354(00)00592-3

UNEP-GEMS(2007). Water Quality Outlook. UNEP-GEMS. Canada.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). Probabilidad y estadística para ingeniería y ciencias (9th ed.). Pearson. https://apunteca.usal.edu.ar/id/eprint/2699/

Wunderlin D.A.., Marıa del Pilar, D., María Valeria, A., Fabiana, P. S., Cecilia, H. A., & María de los Ángeles, B. (2001). Pattern Recognition Techniques for the Evaluation of Spatial and Temporal Variations in Water Quality. A Case Study: Water Research,

Yang, P.-Y., Tsai, J.-T., Chou, J.-H., Ho, W.-H., & Lai, Y.-Y. (2017). Prediction of water quality evaluation for fish ponds of aquaculture. 2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE).

Yildiz, S., & Karakuş, C. B. (2019). Estimation of irrigation water quality index with development of an optimum model: a case study. Environment, Development and Sustainability.

Zare Abyaneh, H. (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. Journal of Environmental Health Science and Engineering, 12(1), 40.