



Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería de Mecánica Eléctrica

**DISEÑO DE INVESTIGACIÓN PARA LA CONSTRUCCIÓN DE MODELOS DE REGRESIÓN Y
SERIES DE TIEMPO DE LA CONCENTRACIÓN DE COLIFORMES FECALES EN EL AGUA
DE LOS RÍOS DE LA CUENCA DEL LAGO DE AMATITLÁN**

Jorge Mario Gutierrez Ovando

Asesorado por el Mtro. Luis Carlos Leonardo Bolaños Méndez

Guatemala, mayo de 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA



FACULTAD DE INGENIERÍA

**DISEÑO DE INVESTIGACIÓN PARA LA CONSTRUCCIÓN DE MODELOS DE REGRESIÓN Y
SERIES DE TIEMPO DE LA CONCENTRACIÓN DE COLIFORMES FECALES EN EL AGUA
DE LOS RÍOS DE LA CUENCA DEL LAGO DE AMATITLÁN**

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA

POR

JORGE MARIO GUTIERREZ OVANDO

ASESORADO POR EL MTRO. LUIS CARLOS LEONARDO BOLAÑOS
MÉNDEZ

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO ELECTRÓNICO

GUATEMALA, MAYO DE 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA



NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Armando Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO


DECANO	Ing. Pedro Antonio Aguilar Polanco
EXAMINADOR	Ing. Helmut Federico Chicol Cabrera
EXAMINADOR	Ing. Walter Giovanni Álvarez Marroquín
EXAMINADOR	Inga. Ingrid Salome Rodríguez de Loukota
SECRETARIA	Inga. Lesbia Magalí Herrera López

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

DISEÑO DE INVESTIGACIÓN PARA LA CONSTRUCCIÓN DE MODELOS DE REGRESIÓN Y SERIES DE TIEMPO DE LA CONCENTRACIÓN DE COLIFORMES FECALES EN EL AGUA DE LOS RÍOS DE LA CUENCA DEL LAGO DE AMATITLÁN

Tema que me fuera asignado por la Dirección de Escuela de Estudios de Postgrado con fecha 31 de enero de 2021.


Jorge Mario Gutierrez Ovando



EEPFI-PP-0483-2022

Guatemala, 31 de enero de 2022

Director
Armando Alonso Rivera Carrillo
Escuela De Ingenieria Mecanica Electrica
Presente.

Estimado Ing. Rivera

Reciba un cordial saludo de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería.

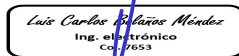
El propósito de la presente es para informarle que se ha revisado y aprobado el Diseño de Investigación titulado: **DISEÑO DE INVESTIGACIÓN PARA LA CONSTRUCCIÓN DE MODELOS DE REGRESIÓN Y SERIES DE TIEMPO DE LA CONCENTRACIÓN DE COLIFORMES FECALES EN EL AGUA DE LOS RÍOS DE LA CUENCA DEL LAGO DE AMATITLÁN**, el cual se enmarca en la línea de investigación: **Todas las áreas - Pronósticos**, presentado por el estudiante **Jorge Mario Gutierrez Ovando** carné número **201213124**, quien optó por la modalidad del "PROCESO DE GRADUACIÓN DE LOS ESTUDIANTES DE LA FACULTAD DE INGENIERÍA OPCIÓN ESTUDIOS DE POSTGRADO". Previo a culminar sus estudios en la Maestría en ARTES en Estadística Aplicada.

Y habiendo cumplido y aprobado con los requisitos establecidos en el normativo de este Proceso de Graduación en el Punto 6.2, aprobado por la Junta Directiva de la Facultad de Ingeniería en el Punto Décimo, Inciso 10.2 del Acta 28-2011 de fecha 19 de septiembre de 2011, firmo y sello la presente para el trámite correspondiente de graduación de Pregrado.

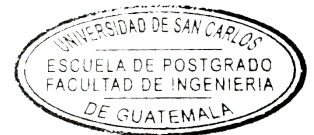
Atentamente,

"Id y Enseñad a Todos"

Mtro. Luis Carlos Bolaños Méndez
Asesor(a)



Mtro. Edwin Adalberto Bracamonte Orozco
Coordinador(a) de Maestría



Mtro. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería





EEP-EIME-0483-2022

El Director de la Escuela De Ingenieria Mecanica Electrica de la Facultad de Ingenieria de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del Asesor, el visto bueno del Coordinador y Director de la Escuela de Estudios de Postgrado, del Diseño de Investigación en la modalidad Estudios de Pregrado y Postgrado titulado: **DISEÑO DE INVESTIGACIÓN PARA LA CONSTRUCCIÓN DE MODELOS DE REGRESIÓN Y SERIES DE TIEMPO DE LA CONCENTRACIÓN DE COLIFORMES FECALES EN EL AGUA DE LOS RÍOS DE LA CUENCA DEL LAGO DE AMATITLÁN**, presentado por el estudiante universitario **Jorge Mario Gutierrez Ovando**, procedo con el Aval del mismo, ya que cumple con los requisitos normados por la Facultad de Ingenieria en esta modalidad.

ID Y ENSEÑAD A TODOS

A handwritten signature in black ink is written over a circular official stamp. The stamp contains the text: "UNIVERSIDAD DE SAN CARLOS DE GUATEMALA", "DIRECCIÓN ESCUELA DE INGENIERIA MECANICA ELECTRICA", and "FACULTAD DE INGENIERIA".

Ing. Armando Alonso Rivera Carrillo
Director
Escuela De Ingenieria Mecanica Electrica

Guatemala, enero de 2022

LNG.DECANATO.OI.359.2022

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería Mecánica Eléctrica, al Trabajo de Graduación titulado: **DISEÑO DE INVESTIGACIÓN PARA LA CONSTRUCCIÓN DE MODELOS DE REGRESIÓN Y SERIES DE TIEMPO DE LA CONCENTRACIÓN DE COLIFORMES FECALES EN EL AGUA DE LOS RÍOS DE LA CUENCA DEL LAGO DE AMATITLÁN**, presentado por: **Jorge Mario Gutierrez Ovando**, después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:



Inga. Aurelia Ariabela Cordova Estrada
Decana

Guatemala, mayo de 2022

AACE/gaoc

ACTO QUE DEDICO A:

- Dios** Por permitirnos estar acá y vivir este momento.
- Mis padres** Roberto Gutierrez y Miriam Ovando, gracias por haber apoyado este sueño y por brindarme su amor a lo largo de toda mi vida.
- Mi hermano** Pablo Gutierrez, mi único hermano, gracias por sus consejos y el apoyo incondicional en cada momento.
- Mi demás familia** Por siempre brindarnos su ayuda, aun en los momentos más difíciles. Gracias de todo corazón.
- Mi novia** Georgina Estrada, gracias por su cariño y constante apoyo en cada uno de mis metas.
- Mis amigos** Luis Aguirre, Pablo Orellana, Kevin Franco, Sergio Herrera; gracias por acompañarme en este viaje lleno de buenos y malos momentos.

AGRADECIMIENTOS A:

**Universidad de San
Carlos de Guatemala**

Por abrirme sus puertas, darme los recursos necesarios para poder crecer profesionalmente y enseñarme sueños colectivos.

Facultad de Ingeniería

Por el conocimiento que adquirí dentro de sus instalaciones.

**Departamento de
matemática**

Por brindarme mi primera experiencia laboral y la oportunidad de ser parte de la formación de futuros ingenieros.

Mi asesor

Por tomarse el tiempo de leer y colaborar con el desarrollo de este trabajo, gracias Ingeniero.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES	V
LISTA DE SÍMBOLOS	VII
GLOSARIO	IX
RESUMEN	XI
1. INTRODUCCIÓN	1
2. ANTECEDENTES	5
3. PLANTEAMIENTO DEL PROBLEMA	11
3.1. Contexto general	11
3.2. Descripción del problema	12
3.3. Formulación del problema	13
3.4. Delimitación del problema	14
4. JUSTIFICACIÓN	15
5. OBJETIVOS	17
5.1. General.....	17
5.2. Específicos	17
6. NECESIDADES POR CUBRIR Y ESQUEMA DE SOLUCIÓN	19
7. MARCO TEÓRICO.....	23
7.1. Fundamentos estadísticos.....	23

7.2.	Análisis de correlación	23
7.2.1.	Correlación de Pearson.....	27
7.2.2.	Correlación de Spearman.....	29
7.2.3.	Análisis de regresión	32
7.2.3.1.	Regresión lineal simple	33
7.2.3.2.	Regresión lineal múltiple	34
7.2.3.3.	Recta de regresión ajustada.....	35
7.2.3.4.	Comparación de modelos.....	37
	7.2.3.4.1. Coeficiente de determinación ajustado.....	37
	7.2.3.4.2. Estadístico CP de Mallows	39
7.3.	Conceptos y definiciones	40
7.3.1.	Lago de Amatlán	40
7.3.1.1.	Ríos de la cuenca.....	40
7.3.2.	Calidad del agua	42
7.3.2.1.	Microorganismos	43
8.	PROPUESTA DE ÍNDICE DE CONTENIDOS	47
9.	METODOLOGÍA	51
9.1.	Características del estudio	51
9.2.	Unidades de análisis	52
9.3.	Variables e indicadores	52
9.4.	Fases del estudio	54
9.4.1.	Fase 1: Revisión de la literatura	54
9.4.2.	Fase 2: Obtención y procesamiento de la información.....	54

9.4.3.	Fase 3: Análisis de regresión.....	54
9.4.4.	Fase 4: Análisis de secuencia temporal	55
9.4.5.	Fase 5: Interpretación de los resultados.....	55
9.4.6.	Fase 6: Elaboración de informe final	55
10.	TÉCNICAS DE ANÁLISIS DE LA INFORMACIÓN	57
10.1.	Técnicas metodológicas	57
10.1.1.	Revisión documental	57
10.1.2.	Meta-análisis.....	57
10.2.	Técnicas estadísticas	57
10.2.1.	Medidas de tendencia central	58
10.2.2.	Medidas de dispersión.....	58
10.2.3.	Análisis gráfico.....	58
10.2.4.	Pruebas de hipótesis	58
10.2.5.	Análisis de residuos.....	59
11.	CRONOGRAMA.....	61
12.	FACTIBILIDAD DEL ESTUDIO	63
12.1.	Recurso humano	63
12.2.	Recursos financieros	63
12.3.	Recursos tecnológicos.....	64
12.3.1.	<i>Software</i>	65
12.3.2.	<i>Hardware</i>	65
12.4.	Acceso a la información.....	65
13.	REFERENCIAS.....	67
14.	APÉNDICE	71

15. ANEXO73

ÍNDICE DE ILUSTRACIONES

FIGURAS

1.	Flujograma del proceso de solución.....	21
2.	Ejemplo gráfico de dispersión y gráfico de cajas	24
3.	Gráfico de cajas con varios niveles.....	25
4.	Relación lineal simple.....	34
5.	Recta ajustada y recta de regresión.....	36
6.	Lago de Amatitlán	41
7.	Cronograma de actividades	61

TABLAS

I.	Interpretación del coeficiente de Pearson	28
II.	Interpretación del coeficiente de Spearman	31
III.	Principales ríos de la cuenca.....	42
IV.	Características de un microorganismo bioindicador.....	44
V.	Operativización de variables	53
VI.	Recursos financieros	64

LISTA DE SÍMBOLOS

Símbolo	Significado
Ca	Calcio
ρ	Coefficiente de correlación de Pearson
R ²	Coefficiente de determinación
ε	Error
°C	Grado centigrado
HCO ₃ ⁻	Ión bicarbonato
Km ²	Kilómetro cuadrado
L	Litro
\bar{x}	Media
pH	Medida de acidez
mg	Miligramo
mL	Mililitro
%	Porcentaje
Q	Quetzales
seg	Segundo
Na ⁺	Sodio
Σ	Sumatoria
→	Tiende a
σ^2	Varianza

GLOSARIO

AIC	Criterio de información de Akaike, es una medida de la calidad relativa de un modelo estadístico.
Autocorrelación	Medida que cuantifica el nivel de asociación que existe entre una variable continua con ella misma.
AMSA	Autoridad para el manejo sustentable de la cuenca del lago de Amatitlán.
ARIMA	Modelo autorregresivo integrado de media móvil.
BIC	Criterio de información de Bayesiano, es una medida de la calidad relativa de un modelo estadístico.
Calidad del agua	Término utilizado para describir las características de una muestra de agua.
Coliforme fecal	Familia de bacterias que se encuentran en grandes cantidades en los excrementos de los seres vivos.
Correlación	Medida que cuantifica el nivel de asociación que existe entre dos variables continuas.
Lago	Cuerpo de agua dulce o salada que se encuentra separada de los océanos.

MAE	Error absoluto medio.
Microorganismo	Seres vivos que se caracterizan por tener un tamaño del orden de los micrómetros y tienen una estructura biológica muy básica.
Regresión multivariada	Análisis que tiene por objetivo estudiar la relación que existe entre las magnitudes de dos o más variables.
RMSE	La raíz cuadrada del error cuadrático medio.
Serie de tiempo	Conjunto de datos ordenados que se encuentran a una distancia equidistante en el tiempo entre sí.

RESUMEN

Este trabajo de graduación propone un diseño de investigación para estudiar la contaminación de los ríos de la cuenca del lago de Amatitlán a través de la cantidad de coliformes fecales presentes en el agua. Los coliformes fecales son microorganismos que se encuentran en los intestinos de los animales de sangre caliente y en las personas, por lo que encontrar un exceso de las mismas en aguas naturales es un indicador de la presencia de contaminación.

La metodología propuesta no se centra en la recopilación de los datos, sino en el procesamiento, transformación e interpretación de los mismos. Para ello se propone partir de la información recabada por la entidad pública encargada del estudio y limpieza de la cuenca del lago de Amatitlán (AMSA). Dicha institución mensualmente realiza un monitoreo sobre los ríos de la cuenca y realiza distintas mediciones químicas, fisicoquímicas y microbiológicas para conocer el estado del agua de los ríos. Se propone dividir el estudio en dos partes, la primera, un estudio transversal donde el objetivo es estudiar la relación que existe entre la cantidad de coliformes fecales presente en los ríos con las medidas químicas y fisicoquímicas de los mismos. Para ello se sugiere utilizar técnicas estadísticas como la regresión lineal simple, la regresión lineal multivariada y la regresión logística. La segunda parte consiste en realizar un estudio del comportamiento de la cantidad de coliformes fecales a través del tiempo, en otras palabras, un estudio longitudinal. Para la segunda parte se sugiere utilizar modelos estadísticos que permitan representar el comportamiento secuencial esperado de la variable, por ejemplo, los modelos de suavizamiento exponencial, los modelos ARIMA, entre otros.

Para finalizar, se presenta un análisis de los recursos que son necesarios para llevar a cabo la investigación. Para el recurso del tiempo se presenta un cronograma con las distintas actividades a realizar. Para el recurso financiero un presupuesto con la cantidad y el costo del material y las herramientas necesarias. Y, por último, una descripción de la cantidad de personas y sus roles para concluir exitosamente la investigación.

1. INTRODUCCIÓN

El presente trabajo tiene por objetivo estudiar la contaminación de los ríos de la cuenca del lago de Amatitlán a través de la estadística multivariada como herramienta de análisis, y a través de los modelos de series temporales como herramientas de predicción. El estudio es una investigación sistemática, ya que el objetivo es aplicar análisis y métodos ya conocidos a un problema en específico.

La contaminación en el lago de Amatitlán se ha incrementado en los últimos años, de sus aguas más del 97 % provienen de los ríos de la cuenca del mismo, por lo que si se quiere reducir es fundamental estudiar el comportamiento de dichos ríos. Por tanto, se requiere, a través de un modelo estadístico, estudiar la relación que existe entre la contaminación de los ríos con el estado de los mismos, y también es importante realizar una proyección de la contaminación de los próximos años.

La importancia de este estudio radica en el hecho que el agua es un componente esencial para la vida de los seres vivos. Por lo que estudiar la contaminación presente en los ríos de la cuenca permitirá crear estrategias efectivas para reducir la misma, y de este modo beneficiar a las comunidades de la región.

Se espera que los modelos construidos durante la investigación permitan relacionar la contaminación fecal de los ríos con sus parámetros ambientales, fisicoquímicos, químicos y también permitan realizar proyecciones sobre los niveles de contaminación para los próximos años. Estos modelos podrán servir

de apoyo al monitoreo mensual realizado por la entidad encargada del manejo sustentable de la cuenca.

La solución propuesta consiste en el cálculo del coeficiente de correlación y la realización de un análisis de regresión multivariado para conocer cómo se relaciona la contaminación de cada río con sus parámetros ambientales, fisicoquímicos y químicos. Una vez realizado el análisis de regresión, se utilizará los valores CP de Mallows y los coeficientes asignados a cada parámetro para identificar cuáles son las variables más influyentes. Identificadas las variables influyentes, se construirán dos modelos de regresión, uno utilizando los parámetros que más información aportan y otro con los parámetros que menor información aportan. Por último, se evaluará la robustez de todos los modelos utilizando los indicadores de información AIC, BIC y el coeficiente de determinación ajustado.

Para la predicción de los niveles de contaminación en los próximos años se realizará un análisis de series de tiempo, con dicho análisis se obtendrán las características de tendencia, estacionalidad, estacionariedad y autocorrelación de los datos. Conociendo las características se construirán los modelos ARIMA y de suavizado exponencial que mejor se ajusten a los datos. Por último, se evaluará la robustez de los modelos construidos por medio de la raíz cuadrada del error cuadrático promedio (RMSE), el error absoluto promedio (MAE) y el coeficiente de información AIC.

El estudio se considera factible debido a que investigaciones previas han obtenido resultados exitosos al emplear estos modelos estadísticos para representar fenómenos ambientales. Es factible económicamente porque la entidad encargada del manejo sustentable de la cuenca ya ha obtenido la información necesaria en los últimos años.

La investigación estará conformada por los siguientes capítulos:

El capítulo 1 incluirá un análisis de los estudios previos más recientes que han servido de guía e inspiración para la investigación, el capítulo 2 tendrá una revisión literaria de las técnicas estadísticas que mejor se ajustan al problema y una revisión a los conceptos fundamentales de la naturaleza del problema, el capítulo 3 presentara los resultados obtenidos del análisis multivariado y el análisis de series de tiempo y el capítulo 4 finalizara con el análisis y discusión de resultados.

2. ANTECEDENTES

La contaminación de un cuerpo de agua es un tema que ha tenido importancia desde hace décadas, desde que se comenzó a estudiar el crecimiento poblacional y el aumento de la urbanización se empezó a cuestionar el impacto que dichos fenómenos sociales generarían en el medio ambiente.

Según Cano (2018):

En su informe, realiza un análisis descriptivo de las mediciones ambientales realizadas en los cuerpos acuíferos de la cuenca del año 2016 al año 2018. De forma visual a través de gráficas de barras y líneas el autor detalla el estado de los ríos durante ese período de tiempo. El autor también realiza un análisis de la contaminación de la cuenca utilizando dos mediciones, las cuales son: el nivel de coliformes fecales existentes dentro del cuerpo de agua y la cantidad de macroinvertebrados encontrados en el agua. Con este estudio se ha podido identificar cuáles son los principales ríos cuenca del lago de Amatitlán. (p. 112)

Gamboa, Becerra, Cifuentes y Rocha (2016):

Realizaron un estudio sobre la contaminación fecal en el embalse La Copa del río Chicamocha en Colombia. Los organismos microbiológicos son uno de los principales bioindicadores utilizados para cuantificar de forma indirecta la calidad del agua. En la investigación se tomaron

muestras de coliformes totales y fecales en 7 puntos distintos del embalse y a lo largo de 6 meses. Entre los resultados obtenidos, se destaca la alta correlación que existe entre los dos tipos de coliformes, por lo que ambos pueden ser utilizados como bioindicadores. También se destaca la independencia de las mediciones a las condiciones ambientales en los 6 meses evaluados. De este estudio se obtuvo la propuesta de utilizar las coliformes fecales como bioindicador. (p. 55)

Una de las funciones de la entidad encargada del manejo sustentable de la cuenca del lago (AMSA), es realizar un monitoreo periódico sobre el estado de los principales ríos de la cuenca. La división de control y calidad ambiental (2020) presenta en sus informes algunos de los parámetros evaluados en dicho monitoreo, entre los cuales podemos mencionar; medidas ambientales como la temperatura y el viento, medidas químicas como la cantidad total de fósforo y nitrógeno, y medidas fisicoquímicas como la demanda de oxígeno. En el documento también se muestra el comportamiento de la contaminación durante el primer año de la pandemia COVID-19. En este estudio se ha podido identificar a AMSA como una posible fuente de información para la presente investigación.

Se realizó un estudio de una proyección del estado ambiental del río Horrood de la provincia de Lorestan en Irán. Los autores implementaron modelos de series temporales para proyectar los siguientes parámetros del agua; TDS, EC, HCO_3^- , SO_4^{2-} , Mg^{2+} , Ca^{2+} , Na^+ y pH. Para evaluar los pronósticos se utilizaron los criterios AIC, R2 y RMSE. Entre los resultados el autor encontró que el modelo ARIMA (autorregresivo, integrado, promedio móvil) fue el que mejor se adaptó a los datos para pronosticar la calidad del agua del río. Los criterios AIC, R2 y RMSE pueden utilizarse para evaluar la

robustez de los modelos de series de tiempo que se construirán en la presente investigación. (Taheri, Ghashaghaie y Georgiou, 2014)

En un estudio de la utilización de series temporales para el pronóstico de parámetros ambientales donde se utilizó como medida de calidad del agua el porcentaje de oxígeno disuelto en el río Vouga Basin en Portugal. El objetivo del estudio fue realizar una comparación de las proyecciones de un modelo ARIMA y un modelo creado a partir de un filtro Kalman, para realizar la comparación entre ambos modelos se utilizó el criterio del coeficiente de determinación. Entre los resultados obtenidos el autor destaca que el modelo ARIMA obtuvo una mejor aproximación en sus pronósticos, sin embargo, la diferencia con respecto al otro algoritmo es leve. El filtro de Kalman podría ser una solución para realizar la proyección de la contaminación en la presente investigación. (Costa, 2015)

Vega Araya y Alvarado Barrantes (2019):

Utilizaron series temporales para estudiar la vegetación de los bosques en la provincia de Guanacaste, Costa Rica. Los autores obtuvieron mediante tecnología de teledetección durante 16 días los valores del índice de área foliar (LAI), el índice normalizado de vegetación (NDVI) y la fracción absorbida de la radiación fotosintéticamente activa (fPAR). Posteriormente como parte de su análisis descompusieron las series de valores en tres componentes; estacionalidad, tendencia y residuos. Entre sus resultados determinaron que los tres parámetros tienen un comportamiento similar, por lo que para futuras investigaciones no es necesario estudiar todos los parámetros. La descomposición de la serie de tiempo en varias componentes es una técnica que puede aplicarse en la presente investigación. (p. 24)

Pisarra (2019):

En su estudio realizó un análisis de regresión lineal multivariado para estudiar la relación entre el ancho y la espesura de la barrera de vegetación que existe entre un arroyo y la actividad humana con la calidad de agua del cuerpo de agua, en esta investigación se utilizó el índice IWQ como medida de calidad del agua. El estudio se llevó a cabo en 8 puntos del área de protección ambiental de la cuenca del río Uberada en Brasil. Entre los resultados obtenidos en el estudio, se destaca que el ancho de la barrera provoca mayor variación en la calidad del agua que la espesura de la barrera. La actividad humana es un factor no considerado que puede proporcionar información sobre la concentración de coliformes fecales en los ríos. (p. 78)

Soo y Seo (2018):

En su estudio se muestran como un mismo problema puede abordarse desde distintos puntos de vista. Los autores discretizaron la concentración de coliformes fecales para clasificar el agua con dos valores posibles agua en buen estado y agua en estado pobre. Luego los autores implementaron un modelo de regresión logística para clasificar muestras de agua según su calidad. El enfoque de discretizar la variable de interés podría aplicarse en la presente investigación. (p. 34)

Zimmer, Brown y Manderson (2018):

Estudiaron los excedentes de coliformes fecales producidos por mariscos en la Bahía Tillamook, Oregón, USA. Las coliformes fecales no siempre

están asociados a desechos humanos. Los autores construyeron distintos modelos utilizando las técnicas de; regresión multivariada, regresión logística, arboles de decisión y regresión de efectos mixtos. Los modelos relacionaron la precipitación y la altura de la marea con el excedente de coliformes fecales permitido. Entre los resultados obtenidos destaca que el árbol de decisión y la regresión logística fueron los modelos que obtuvieron mejores predicciones. Este estudio muestra una fuente de contaminación no tomada en cuenta hasta ahora, la generada por los mismos seres vivos que viven en el agua. (p. 55)

Morantes, Polo y Pérez (2019):

En su investigación realizada dejando por un lado la contaminación del agua, se utilizó la estadística multivariada para estudiar la contaminación atmosférica del campus de la Universidad Simón Bolívar, Caracas, Venezuela. Se utilizaron 9 variables independientes para estimar la cantidad de material particulado (PM) en el aire. Para evaluar la robustez del modelo se utilizó el coeficiente de determinación y el error de sesgo promedio. La métrica de error del sesgo promedio puede utilizarse para evaluar la robustez de los modelos de regresión propuestos en la presente investigación. (p. 11)

Como se ha podido observar en la revisión anterior, la estadística es una herramienta muy utilizada en otros países para el estudio de los cuerpos naturales de agua, por lo que será muy útil en el análisis y predicción de la contaminación de la cuenca del lago de Amatitlán.

3. PLANTEAMIENTO DEL PROBLEMA

3.1. Contexto general

El lago de Amatitlán es un lago de origen volcánico, es el quinto lago más grande de Guatemala con un área de espejo promedio de 15,2 Km² aproximadamente, es posiblemente el depósito natural de agua más contaminado del país, contaminación que se ha ido incrementando en los últimos años. (Cano, 2018, p.11)

En respuesta a los altos índices de contaminación del lago, el Congreso de la República de Guatemala creó por medio del decreto 64-96 la entidad encargada del manejo sustentable de la cuenca del lago de Amatitlán (AMSA), entre las atribuciones de esta institución están planear, organizar y ejecutar las tareas necesarias para disminuir la contaminación y recuperar el ecosistema de la cuenca del lago. “Como parte de sus funciones la entidad ha mantenido un monitoreo periódico sobre puntos estratégicos, entre los cuales se puede mencionar; la playa pública, centro este, centro oeste, laguna de calderas y los principales ríos de la cuenca” (Congreso de la República de Guatemala, 1996, p. 15).

Más del 97 % del agua del lago proviene de los ríos que desembocan en el mismo, por lo que estudiar la contaminación de los ríos y como esta se relaciona con el estado de los mismos es fundamental para la recuperación del lago de Amatitlán. (División de Control y Calidad, 2020)

3.2. Descripción del problema

La cuenca del lago de Amatitlán se compone de 18 ríos, de los cuales únicamente 2 desembocan directamente en las aguas del lago, siendo los ríos Villalobos y Pampumay, los demás ríos se unen al cauce de los dos anteriores.

El río Villalobos aporta el 95 % del agua del lago, mientras que el río Pampumay aporta solo el 2,71 %. Como parte del monitoreo AMSA realizo mediciones periódicas de distintos parámetros fisicoquímicos, químicos y microbiológicos en 7 de los ríos de la cuenca en el período 2016 - 2021. (Cano, 2018)

Entre las mediciones realizadas por AMSA, destaca el análisis microbiológico de coliformes fecales.

Las coliformes fecales son bacterias que se encuentran en grandes cantidades en los intestinos de las personas y en los animales de sangre caliente, por lo que es muy común encontrarlas en aguas residuales provenientes de ciudades, granjas, entre otros. La cantidad de estas bacterias presentes en el agua es un bioindicador ampliamente utilizado para medir la contaminación del agua, ya que la cantidad de colonias de estas bacterias presentes en una muestra de agua de 100 ml es directamente proporcional al nivel de contaminación. (Fernández Santisteban, 2017)

Por lo expuesto anteriormente, es necesario estudiar, a través de un modelo estadístico, las relaciones que existen entre la concentración de coliformes fecales de los ríos de la cuenca y sus características fisicoquímicas y químicas. Así como también es importante realizar una proyección de la cantidad de coliformes fecales de los ríos en los próximos años.

3.3. Formulación del problema

A continuación se muestra la pregunta central del problema y las preguntas auxiliares.

- Pregunta central
 - ¿Cuáles son los modelos estadísticos que mejor se ajustan a las mediciones realizadas por AMSA del año 2016 al año 2021, para relacionar los factores fisicoquímicos y químicos con la concentración de coliformes fecales, y para realizar predicciones sobre la cantidad de coliformes fecales de los ríos de la cuenca del lago de Amatitlán? (AMSA, 2021)

- Preguntas auxiliares

Para responder a esta interrogante se deberán contestar las siguientes preguntas auxiliares:

- ¿Cuál es el grado de correlación que existe entre el nivel de coliformes fecales de los ríos de la cuenca con sus factores fisicoquímicos y químicos?

- ¿Qué variables fisicoquímicas y químicas aportan más información sobre la concentración de coliformes fecales en los ríos de la cuenca?

- ¿Cuál es la robustez de los modelos de regresión que relacionan la concentración de coliformes fecales con las variables fisicoquímicas y químicas de los ríos de la cuenca?
- ¿Cuál será el nivel de coliformes fecales de los principales ríos de la cuenca del lago de Amatitlán en los años 2022 - 2023?

3.4. Delimitación del problema

Para tratar este problema se utilizará la información obtenida mensualmente del monitoreo que realizó AMSA desde el año 2016 hasta el año 2021 de los siguientes ríos de la cuenca; Pampumay, Frutal, Pinula, Platanitos, San Lucas, y el río Villalobos.

4. JUSTIFICACIÓN

La presente investigación se desarrolla en el campo de la ingeniería ambiental, específicamente en la aplicación de la estadística multivariada y la realización de pronósticos a la calidad del agua. El objetivo principal es estudiar la contaminación de los principales ríos de la cuenca del lago de Amatitlán mediante el bioindicador de coliformes fecales.

La importancia de la investigación radica en el hecho de que la cuenca del lago de Amatitlán posee los cuerpos de agua más contaminados de Guatemala, por lo que los resultados de las investigaciones realizadas en ella pueden utilizarse como una guía para futuras investigaciones en los demás cuerpos de agua del país.

Según estudios anteriores un alto porcentaje de las aguas del lago de Amatitlán proviene de los ríos que pertenecen a su cuenca, por lo que para resolver el problema de la contaminación en el lago es fundamental comprender la contaminación de los ríos primero.

Como resultado de la investigación se tendrán modelos estadísticos que permitirán analizar el comportamiento de la contaminación de los ríos a través del tiempo y permitirán realizar proyecciones de la misma para los próximos años. También se tendrá un modelo que permitirá relacionar la concentración de coliformes fecales con los parámetros ambientales, fisicoquímicos y químicos de los ríos.

Los modelos podrán ser utilizados por la entidad encargada del manejo sustentable de la cuenca (AMSA) en su investigación para reducir la cantidad de contaminación presente en los ríos, por consecuente, también beneficiara a las comunidades aledañas que utilizan el agua de los mismos en sus actividades diarias.

5. OBJETIVOS

5.1. General

Construir un modelo de regresión que relacione el estado fisicoquímico del agua de los ríos de la cuenca del lago de Amatitlán con su concentración de coliformes fecales, y un modelo de series de tiempo para realizar proyecciones de la cantidad de coliformes fecales de los mismos.

5.2. Específicos

- Estimar el nivel de correlación entre la concentración de coliformes fecales y los factores fisicoquímicos y químicos a través del coeficiente de Pearson, para identificar el tipo de relación de cada uno de los factores.
- Identificar las variables fisicoquímicas y químicas que más información aportan sobre la concentración de coliformes fecales mediante un análisis de regresión, para descartar las variables insignificantes que no aportan información.
- Calcular la robustez de los modelos de regresión que relacionan la concentración de coliformes fecales con las variables fisicoquímicas y químicas mediante los indicadores AIC, BIC y R2 ajustado. Para identificar cual es el modelo que mejor se ajusta a los datos.

- Estimar el nivel de coliformes fecales de los ríos de la cuenca para el año 2022 y año 2023 mediante la construcción de un modelo ARIMA de series de tiempo, para conocer cuál será el estado de los ríos en los próximos años.

6. NECESIDADES POR CUBRIR Y ESQUEMA DE SOLUCIÓN

Se desconoce la relación que existe entre la cantidad de coliformes fecales presentes en el agua (contaminación) con los parámetros fisicoquímicos y químicos de la misma. A su vez, también se desconoce que parámetros son los que pueden proporcionar más información sobre la presencia de dichas bacterias en el agua. Por último, se desconoce cuáles serán los niveles de contaminación que tendrán los ríos de la cuenca en los próximos años.

Primero, se realizará una solicitud a la entidad AMSA para tener acceso a la información del monitoreo mensual que realiza la institución sobre los ríos de la cuenca del lago de Amatitlán.

Luego se calculará el coeficiente de correlación de Pearson entre cada uno de los factores fisicoquímicos y químicos con la cantidad de coliformes fecales presentes en el agua, de este modo se podrá cuantificar la intensidad y el tipo de relación (positiva o negativa) que existe entre los parámetros con la contaminación.

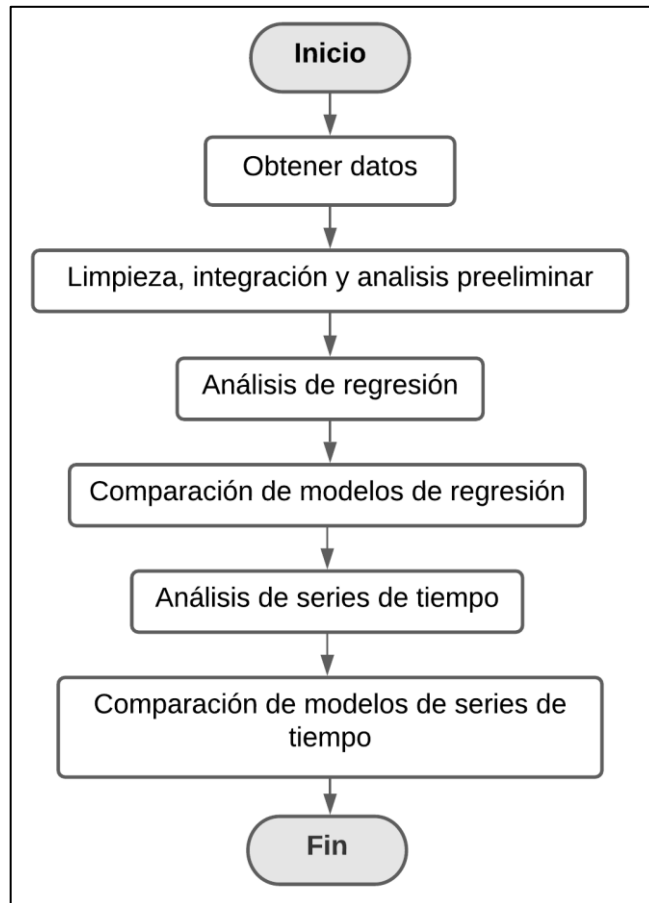
Se realizará un análisis de regresión con los parámetros ambientales, fisicoquímicos y químicos como las variables independientes, y la concentración de coliformes fecales presente en el agua como la variable dependiente. Luego se identificarán los parámetros que más influencia tienen o que más información aportan sobre la variable de interés en el modelo de regresión, para ello se utilizarán los valores CP de Mallows, el intervalo y la magnitud de los coeficientes de cada variable.

Conociendo que variables aportan más información y que variables aportan menos, se construirá un modelo de regresión utilizando cada conjunto de variables independientes.

Después se realizará una comparación entre los modelos de regresión construidos y se seleccionará el modelo que tenga mayor robustez, para ello se utilizará el coeficiente de determinación ajustado, y los indicadores de información AIC y BIC.

Por último, para conocer cuál será la cantidad de coliformes fecales presente en los ríos en los próximos años, se construirá un modelo ARIMA de series de tiempo para cada uno de los ríos. Con estos modelos se podrá expresar la naturaleza secuencial de la contaminación como una función lineal que dependerá de los valores anteriores y de los errores aleatorios. Los modelos ARIMA se construirán minimizando la raíz cuadrada del error cuadrático medio (RMSE) que existe entre la predicción de los modelos y los valores reales.

Figura 1. **Flujograma del proceso de solución**



Fuente: elaboración propia.

7. MARCO TEÓRICO

7.1. Fundamentos estadísticos

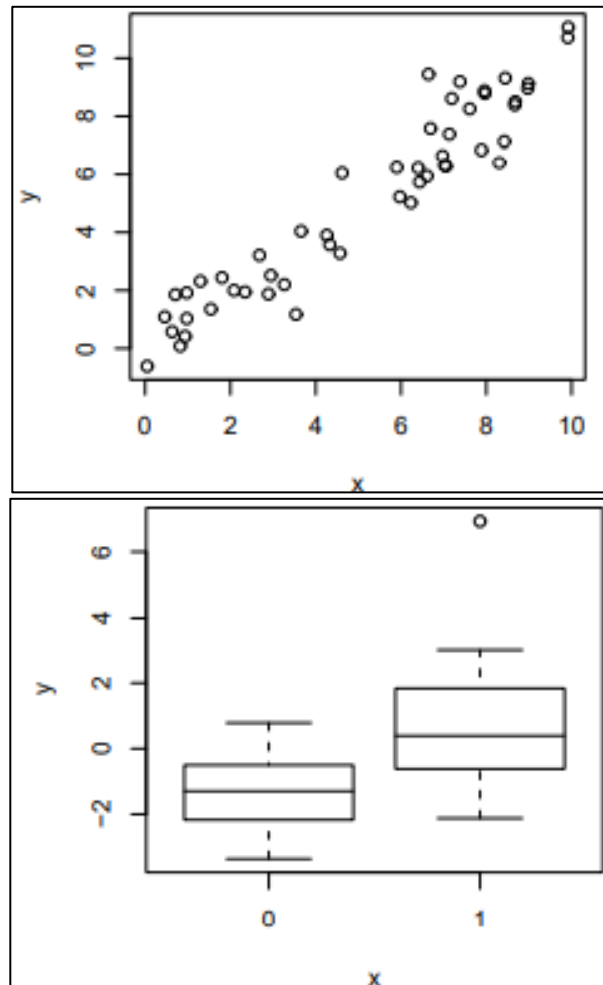
En esta sección se hará una revisión documental de los conceptos estadísticos que servirán de base para fundamentar la aplicación de la estadística en el desarrollo de la investigación.

7.2. Análisis de correlación

“El análisis de correlación se utiliza cuando se desea conocer la intensidad y el tipo de asociación que existe entre dos variables. Comúnmente las dos variables representan muestras de dos poblaciones distintas” (Navidi, 2006, p. 78).

“El análisis de asociación se puede realizar de forma visual utilizando gráficos de dispersión cuando ambas variables son continuas, o utilizando gráficos de cajas cuando una de las variables es continua y la otra es categórica con distintos niveles” (Aparicio, Martínez y Morales, 2004 p. 55).

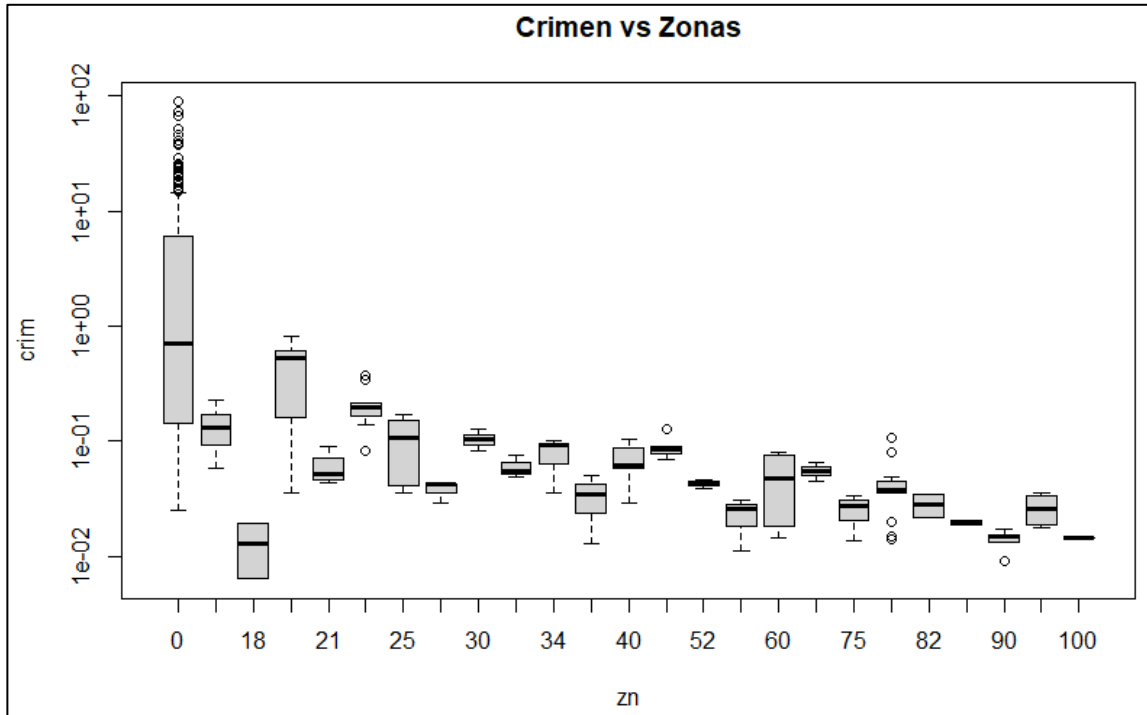
Figura 2. **Ejemplo gráfico de dispersión y gráfico de cajas**



Fuente: elaboración propia, utilizando el lenguaje de programación R.

En los ejemplos mostrados en la figura anterior, ambas graficas muestran una asociación directa entre el par de variables (x,y). Cuando la variable categórica del grafico de cajas tiene demasiados niveles se vuelve complicada la interpretación del mismo, por ejemplo, en la siguiente figura la comparación del crimen entre las distintas zonas se complica entre más zonas tenga la ciudad.

Figura 3. Gráfico de cajas con varios niveles



Fuente: Aparicio, Martínez y Morales (2004.) *Modelos lineales aplicados en R*.

La inspección visual puede proporcionar resultados erróneos debido a que depende en gran medida de la perspicacia y el criterio del observador de los gráficos.

Para evitar ese problema se creó el coeficiente de correlación, el cual es una medida numérica que permite cuantificar la asociación entre dos variables. “Es importante aclarar que el coeficiente de correlación únicamente se puede utilizar cuando ambas variables son numéricas” (Navidi, 2006, p. 6).

Se le llama coeficiente de correlación poblacional cuando se utilizan todos los datos de una o más poblaciones para calcular el mismo. Por otro

lado, se le llama coeficiente de correlación muestral cuando los datos provienen de muestras aleatorias de una o más poblaciones, que es lo que ocurre en la mayoría de las veces. Con el coeficiente de correlación muestral es posible crear intervalos de confianza y evaluar pruebas de hipótesis sobre las poblaciones, es decir, es posible realizar estimaciones sobre el comportamiento que tienen todos los datos poblacionales. (Navidi, 2006)

Aparicio, Martínez y morales (2004):

Explican que los valores que puede adquirir el coeficiente de correlación están dentro del rango $[-1,1]$. Si el valor del coeficiente es positivo significa que la relación es directamente proporcional, es decir, si una de las variables aumenta su valor la otra también aumenta. Por el contrario, si el valor es negativo significa que la relación es inversamente proporcional, en otras palabras, si una de las variables aumenta la otra disminuye. Por último, en el caso que el valor del coeficiente sea igual a 0, significa que no existe ningún tipo de asociación entre ambas variables. (p. 11)

Algo a tener muy en cuenta cuando se trabaja con correlaciones es el fenómeno de confusión.

Este fenómeno de confusión ocurre cuando dos variables se encuentran correlacionadas debido a la existencia de una tercera variable que está relacionada con ambas variables. Por ejemplo, si se considera estudiar la altura de los niños con la cantidad de palabras que conocen de un idioma en específico, al analizar la asociación entre ambas variables se puede encontrar que existe una relación positiva, entre más altura tiene el niño más palabras

conoce, sin embargo, esa relación existe porque ambas variables están relacionadas con una tercera que es la edad del niño. (Navidi, 2006)

Caycho, Castillo y Merino (2020):

Aunque el coeficiente de Pearson es el más ampliamente conocido, existen otros coeficientes como el de Kendall y el de Spearman. El coeficiente de Pearson únicamente se puede utilizar cuando se tienen pruebas suficientes que los datos tienen el comportamiento de una distribución normal, en caso contrario, lo más adecuado es utilizar un coeficiente que no se vea afectado por la falta de normalidad en los datos, como los coeficientes de correlación de Kendall o el de Spearman. (p. 16)

7.2.1. Correlación de Pearson

Aparicio, Martínez y Morales definen el coeficiente de correlación de Pearson para dos poblaciones (X, Y) de la siguiente manera:

$$\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)*\text{Var}(Y)}} \quad (\text{Ec. 1})$$

Donde:

Cov(X,Y) = representa la covarianza entre ambas poblaciones

Var(X) = es la varianza de la población X

Var(Y) = es la varianza de la población Y

Para el caso de dos variables (x,y), que representan datos muestrales de dos poblaciones se utiliza la siguiente expresión.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (\text{Ec. 2})$$

Donde:

x_i = representa el elemento i del conjunto de datos x

y_i = es el elemento i del conjunto de datos y

\bar{x} = representa la media del conjunto de datos x

\bar{y} = es la media del conjunto de datos y

Con la magnitud del coeficiente resultante se puede interpretar como es la intensidad y el tipo de asociación que existe entre los dos conjuntos de datos. En la siguiente tabla se muestra un resumen de las interpretaciones que se pueden realizar en base a la magnitud.

Tabla I. **Interpretación del coeficiente de Pearson**

Magnitud	Categoría
$ \rho = 0$	Ausencia de correlación
$ \rho \rightarrow 0$	Correlación débil
$ \rho \rightarrow 1$	Correlación fuerte
$ \rho = 1$	Correlación perfecta

Fuente: elaboración propia.

7.2.2. Correlación de Spearman

Conocido como el coeficiente de correlación de rangos de Spearman, se caracteriza porque no es necesario que los datos tengan una distribución normal, tampoco que sean datos cuantitativos, por lo que se pueden utilizar variables cualitativas ordinales.

“La diferencia principal de este coeficiente es que no trabaja con los valores puntuales de las variables, sino que trabaja con los rangos de las mismas” (Caycho, Castillo y Merino, 2020, p. 33).

Para calcular el coeficiente se utiliza el siguiente procedimiento:

- Se ordenan los valores de cada variable del menor valor al mayor valor, después se le asigna a cada valor un rango correspondiente.
- Para el caso que se tengan valores repetidos, para cada valor repetido se calcula la cantidad de $t^3 - t$, donde t es la cantidad de veces que se repite el valor. Luego se suman las cantidades calculadas para cada variable, dichas sumas reciben los nombres de ST_x y ST_y , respectivamente.
- Se procede a calcular la diferencia que existe entre los valores de los rangos de ambos conjuntos.

$$d_i = R_{xi} - R_{yi} \quad (\text{Ec. 3})$$

Donde:

d_i = es la i diferencia

R_{xi} = es el rango del i elemento del conjunto x

R_{yi} = es el rango del i elemento del conjunto y

- Por último, se calcula el coeficiente de correlación con la siguiente expresión:

$$\rho = \frac{T_x + T_y - \sum_{i=0}^n d_i^2}{2\sqrt{T_x T_y}} \quad (\text{Ec. 4})$$

$$T_x = \frac{n^3 - n - ST_x}{12} \quad (\text{Ec. 5})$$

$$T_y = \frac{n^3 - n - ST_y}{12} \quad (\text{Ec. 6})$$

Donde:

n = es la cantidad de elementos en cada conjunto de datos

d_i = es la diferencia del rango entre los i elementos de cada conjunto

ST_x y ST_y = son los ajustes por valores repetidos de los conjuntos x y y

En la tabla II se muestra la interpretación de la magnitud del coeficiente de correlación por rangos de Spearman.

Tabla II. Interpretación del coeficiente de Spearman

Magnitud	Categoría
$\rho > 0$	Existe concordancia
$\rho = 1$	Concordancia perfecta
$\rho < 0$	Existe discordancia
$\rho = -1$	Discordancia perfecta
$\rho = 0$	No hay discordancia ni concordancia

Fuente: elaboración propia.

También conocido como coeficiente de correlación por rangos de Kendall, al igual que el coeficiente de Spearman mide el grado e intensidad de la relación que hay entre los rangos de los valores. Por lo que tampoco es necesario que los datos tengan una distribución muestral y que sean cuantitativos.

“El coeficiente τ se define como la diferencia entre las probabilidades de encontrar concordancia y discordancia entre los conjuntos de datos” (Caycho, Castillo y Merino, 2020, p. 36).

Existe concordancia entre dos observaciones (x_i, y_i) con otro par (x_{i+1}, y_{i+1}) sí al realizar las restas $x_{i+1} - x_i$ y $y_{i+1} - y_i$ se obtiene el mismo signo. Por otro lado, si se obtiene un signo distinto se dice que las observaciones son discordantes. Si $x_{i+1} = x_i$ o $y_{i+1} = y_i$ no existe concordancia ni discordancia, en este caso se dice que las observaciones son coincidentes.

Si hay una mayor cantidad de pares de observaciones concordantes se dice que la asociación entre las variables es positiva. Por el contrario, si la mayoría de pares de observaciones son discordantes la asociación es negativa. En el caso que la cantidad de pares discordantes sea igual a la cantidad de

pares concordantes, se dice que no existe asociación alguna entre las variables. Para calcular el coeficiente de Kendall se utiliza la siguiente expresión:

$$\tau = \frac{P-Q}{\sqrt{\frac{n(n-1)}{2}-T_x}\sqrt{\frac{n(n-1)}{2}-T_y}} \quad (\text{Ec. 7})$$

$$T_x = \frac{1}{2} \sum t_x(t_x - 1) \quad (\text{Ec. 8})$$

$$T_y = \frac{1}{2} \sum t_y(t_y - 1) \quad (\text{Ec. 9})$$

Donde:

P = cantidad de pares concordantes

Q = cantidad de pares discordantes

n = cantidad de pares de observaciones

T_x = es el ajuste por rangos repetidos en el conjunto de datos x

T_y = es el ajuste por rangos repetidos en el conjunto de datos y

t_x = es el número de observaciones en el conjunto x que están empatadas en un rango determinado

t_y = es el número de observaciones en el conjunto y que están empatadas en un rango determinado

7.2.3. Análisis de regresión

Es utilizado en problemas en los que se conoce que puede existir una relación entre dos variables o más variables. Por un lado, se tienen las variables dependientes, o de respuesta, estas son las variables de interés en el

problema que se desea estudiar, y por otro lado están las variables independientes, las cuales generan cambios o influyen en el comportamiento de las variables dependientes.

La relación entre dos variables puede ser de dos formas; determinista o probabilística. Cuando la relación es determinista, un valor de x produce siempre el mismo valor de y . “Por otro lado, cuando la relación es probabilística un mismo valor de x puede generar distintos valores de y . Las relaciones que se estudiarán en la investigación corresponden a las relaciones probabilísticas” (Walpole, Myers y Ye, 2012, p. 40).

7.2.3.1. Regresión lineal simple

“La variable independiente y la variable dependiente se pueden relacionar de distintas maneras, siendo la más simple la relación lineal. definen la relación lineal de dos variables de la siguiente manera” (Walpole, Myers y Ye, 2012, p. 42).

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (\text{Ec. 10})$$

Donde:

Y = es la variable dependiente

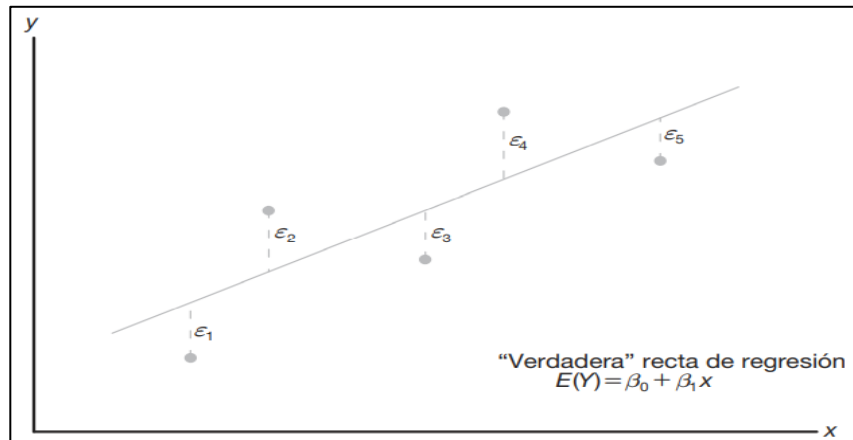
X = es la variable independiente

β_0 = es la intersección

β_1 = es la pendiente

ε = representa la variación aleatoria de la relación con $E(\varepsilon) = 0$ y varianza constante σ^2

Figura 4. **Relación lineal simple**



Fuente: Walpole, Myers y Ye, (2012). *Probabilidad y estadística para ingeniería y ciencias*.

En la gráfica anterior se puede observar claramente como los datos no coinciden siempre con la recta que representa la relación entre las variables, dicha variación o error es causada por el componente aleatorio. Como el valor medio de dicho error es igual a cero, los datos (x_i, y_i) estarán dispersos alrededor de la recta con una variación constante (σ^2) .

7.2.3.2. **Regresión lineal múltiple**

La mayoría de los problemas de la vida real son tan complejos que una única variable independiente no es suficiente para lograr explicar el comportamiento de la variable dependiente. "Se denominan regresión múltiple al caso cuando se tienen dos o más variables independientes." (Walpole, Myers y Ye, 2012, p. 45)

La regresión múltiple se define de la siguiente manera:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon \quad (\text{Ec. 11})$$

Donde:

Y = es la variable dependiente

X_1 = es la primera variable independiente

X_n = es la n variable independiente

β_0 = es la intersección

β_1 = es el coeficiente asociado a la primera variable independiente

β_n = es el coeficiente asociado a la n variable independiente

ε = representa la variación aleatoria de la relación con $E(\varepsilon) = 0$ y varianza constante σ^2

7.2.3.3. Recta de regresión ajustada

El objetivo principal de un análisis de regresión es obtener información sobre los parámetros beta ($\beta_0, \beta_1, \dots, \beta_n$).

Según Novales (2010):

Hace la aclaración que es común encontrarlos en la literatura como coeficientes de correlación. Para obtener los valores exactos de los coeficientes se debe conocer el conjunto de todos los valores posibles que pueden tener las variables Y, X_1, X_2, \dots, X_n , en otras palabras, conocer a todas las poblaciones. Como en la práctica es imposible tener todos los valores poblacionales, se utilizan datos muestrales representados como y, x_1, x_2, \dots, x_n . (p. 90)

Por lo que la expresión de la regresión múltiple con datos muestrales es:

$$y = b_0 + b_1x_1 + \dots b_nX_n + \varepsilon \quad (\text{Ec. 12})$$

Donde:

y = son datos muestrales de la variable dependiente

x_1, x_n = son datos muestrales de las variables independientes

b_0 = es la intersección

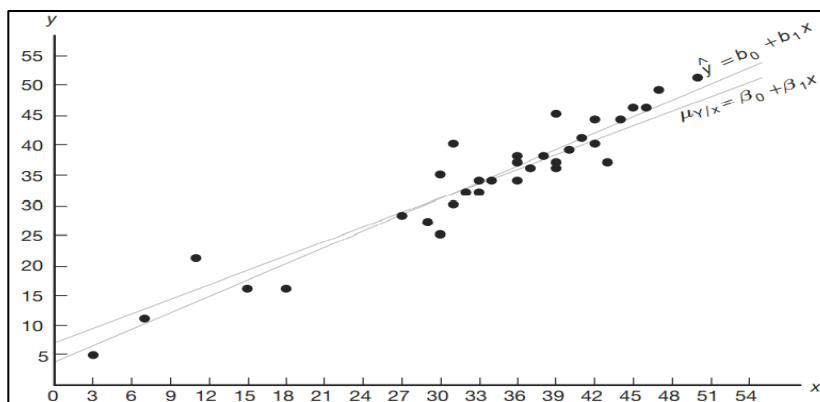
b_1 = es el coeficiente asociado a la primera variable independiente

b_n = es el coeficiente asociado a la n variable independiente

ε = representa el error aleatorio

Los coeficientes $(b_0, b_1, \dots b_n)$ obtenidos con datos muestrales son estimaciones de los coeficientes de regresión poblacionales $(\beta_0, \beta_1, \dots \beta_n)$, es decir, la recta (Ec. 12) denominada como recta ajustada es solo una aproximación de la recta de regresión (Ec. 11) (Novales, 2010).

Figura 5. **Recta ajustada y recta de regresión**



Fuente: Walpole, Myers y Ye, (2012). *Probabilidad y estadística para ingeniería y ciencias*.

7.2.3.4. Comparación de modelos

Como se mencionó anteriormente, la relación entre dos o más variables puede ser lineal, polinomial, logarítmica, entre otros. Cada tipo de relación se puede expresar con un modelo distinto, por lo que es importante cuantificar que tanto se ajustan los modelos a los datos.

“Denominan a dicho ajuste como robustez del modelo, entre mayor sea la robustez del modelo mayor capacidad tendrá el mismo para explicar el comportamiento de los datos.” (Aparicio, Martínez y Morales, 2004, p. 21)

Para evaluar la robustez existen distintos criterios, entre los cuales se pueden mencionar los siguientes:

- Coeficiente de determinación ajustado
- El estadístico CP de Mallows
- El criterio de información de Akaike (p. 122)

7.2.3.4.1. Coeficiente de determinación ajustado

El coeficiente de determinación es una medida que permite cuantificar la cantidad de variación de los datos que logra ser explicada por un modelo de regresión. Walpole, Myers y Ye (2012) lo definen de la siguiente manera:

$$R^2 = \frac{\sum_{i=1}^m (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^m (y_i - \bar{y})^2} \quad (\text{Ec. 13})$$

Donde:

y_i = es el dato i de la variable dependiente

\hat{y}_i = es la predicción i de la variable dependiente

\bar{y} = es la media de los variables de la variable dependiente

m = es la cantidad de datos muestrales utilizados para la construcción del modelo

Uno de los mayores problemas que tiene el coeficiente de determinación es que no toma en cuenta la complejidad del modelo, es decir, la cantidad de variables independientes que posee el mismo, siempre se debe buscar seleccionar el modelo con la menor cantidad de variables independientes posible, en otras palabras, el modelo con la menor complejidad. Para solucionar el inconveniente al coeficiente de determinación se le agrega el siguiente ajuste. (Aparicio, Martínez y Morales, 2004)

$$R_a^2 = 1 - \frac{m-1}{m-n} (1 - R^2) \quad (\text{Ec. 14})$$

Donde:

R^2 = es el coeficiente de determinación normal

n = es la cantidad de variables independientes utilizadas en la construcción del modelo

m = es la cantidad de datos muestrales utilizados para la construcción del modelo

“Se denomina a este coeficiente como el coeficiente de determinación ajustado. A diferencia del coeficiente normal, la versión ajustada penaliza la

cantidad de variables independientes que tenga el modelo.” (Aparicio, Martínez y Morales, 2004, p. 22)

7.2.3.4.2. Estadístico CP de Mallows

Este criterio se utiliza para comparar modelos similares que poseen una cantidad distinta de variables independientes. “Se basa en el cálculo de la suma del error cuadrático que tienen los predichos generados por el modelo. El criterio se calcula de la siguiente manera” (Aparicio, Martínez y Morales, 2004, p. 23).

$$C_p = \frac{SSE_P}{MSE} - (m - 2n) \quad (\text{Ec. 15})$$

$$SSE_P = \sum_{i=0}^m (y_i - \hat{y}_i)^2 \quad (\text{Ec. 16})$$

Donde:

y_i = es el i valor de la variable dependiente

\hat{y}_i = es el i predicho del modelo creado usando únicamente n variables independientes

n = es el número de variables independientes que utiliza el modelo

m = es la cantidad de datos muestrales utilizados para la construcción del modelo

MSE = es el error del modelo utilizando todas las variables independientes disponibles

El criterio para seleccionar un modelo utilizando el estadístico C_p de Mallows, es seleccionar al modelo que proporcione un valor de C_p que más se parezca a p , y también el que tenga la menor magnitud.

7.3. Conceptos y definiciones

En esta sección se hará una revisión documental de los conceptos clave relacionados a la calidad del agua, los cuales servirán de base para comprender el problema de investigación.

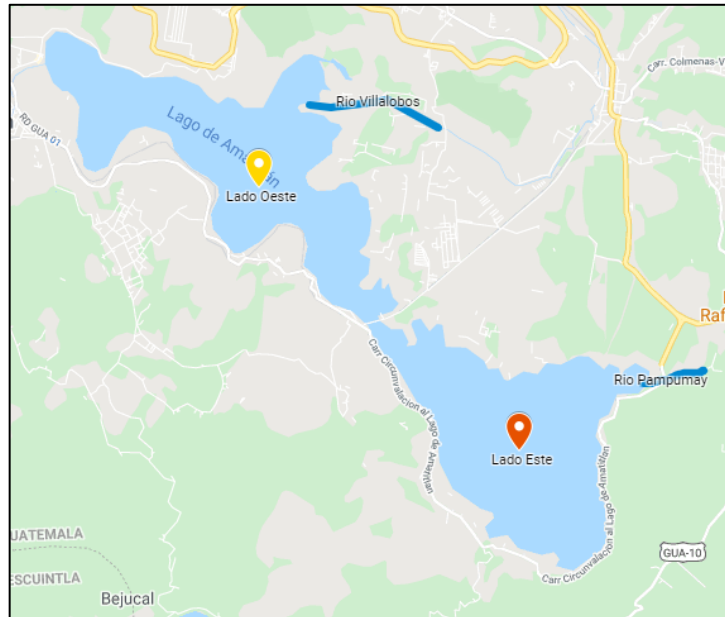
7.3.1. Lago de Amatitlán

El lago de Amatitlán está ubicado en la región sur del país, específicamente en el departamento de Guatemala. Está catalogado como un lago de tipo urbano debido a la cantidad de áreas urbanas que hay en sus alrededores. Como todo cuerpo de agua natural, su profundidad cambia en función de la época del año y el efecto del clima, indica que en promedio el área espejo del lago a lo largo del año es de 15.2 Km^2 . (Cano, 2018)

7.3.1.1. Ríos de la cuenca

A la región geográfica que involucra el cuerpo del lago y sus nacimientos de agua se le denomina como cuenca del lago de Amatitlán, que incluye decenas de comunidades, 18 ríos y la laguna de Calderas. Los dos ríos más importantes son el río Villalobos y el río Pampumay, ya que son los únicos ríos que desembocan directamente en el cuerpo del lago, se identificó que el 95 % de las aguas del lago provienen del primero, mientras que del segundo provienen el 2,71 % de las aguas, por último, el agua restante se estima que proviene de las temporadas de lluvia a lo largo del año. (Cano, 2018)

Figura 6. **Lago de Amatlán**



Fuente: Google Maps. *Mapa del lago de Amatlán*. Consultado el 24 de octubre de 2021.
Recuperado de <https://mymaps.google.com>.

“El lago de Amatlán posee en su zona más estrecha un relleno ferroviario construido a finales del siglo XIX, dicha construcción divide el lago en dos regiones (Este y Oeste), y limita el cauce entre ambas” (División de Control y Calidad, 2020, p. 7).

Pérez (2007):

Indica en su investigación que el flujo natural del agua de la región Este se vio afectado por la construcción del relleno, debido a que la única salida de las aguas del lago es el río Michatoya que se ubica en la región Este. (p. 77)

Los ríos más influyentes de la cuenca del lago que aportan la mayor cantidad de agua al cuerpo del lago son:

Tabla III. **Principales ríos de la cuenca**

Nombre
Río El Frutal
Río Pampumay
Río Pansalic
Río Pinula
Río Platanitos
Río San Lucas
Río Villalobos

Fuente: elaboración propia.

7.3.2. Calidad del agua

La calidad del agua es un término muy complejo de medir, ya que es muy subjetivo y depende en gran medida de la perspectiva de la persona que realizara la medición. Por ejemplo, la calidad de agua no será la misma para un agricultor que para una persona que lava carros, mientras el primero está interesado en que el agua no tenga contaminantes químicos que dañen sus cultivos, el segundo querrá que el agua se encuentre clara sin desechos sólidos.

La calidad del agua puede definirse en función del uso que se le dé a la misma o en base a alguna propiedad de interés. Por ejemplo, en el caso anterior el agricultor puede medir la calidad del agua en función de la cantidad de contaminantes disueltos en la misma, y el lavador de carros en base a la transparencia del agua. (Dirección General de Obras Hidráulicas, 2000)

Pérez, Nardini y Galindo (2018):

Denominan a estas propiedades que están relacionadas con la calidad del agua de manera indirecta y que permiten cuantificarla como métricas de calidad. Para los cuerpos de agua natural se suele utilizar a los seres vivos presentes en el agua como un indicador de calidad, por ejemplo, hay investigaciones donde han utilizado a los peces, las algas, los insectos, bacterias, entre otros. El bioindicador que se utilizara en el estudio para cuantificar los niveles de contaminación de los ríos son los microorganismos. (p. 43)

7.3.2.1. Microorganismos

Los microorganismos son los bioindicadores más utilizados para evaluar los niveles de contaminación presentes en un cuerpo de agua.

Bautista (2013):

“Destaca entre sus características su capacidad para sobrevivir en entornos altamente contaminados, su reducido tamaño (imperceptibles al ojo humano), por último, su existencia en todas partes” (p. 15).

Realizar un análisis profundo de todos los microorganismos presentes en el agua es una tarea ardua y costosa, por lo que el análisis se suele centrar en una determinada familia o subgrupo de microorganismos.

Diaz, Fall y Jimenez (2003):

Explican que cada familia se comporta de manera diferente frente a las condiciones ambientales de la naturaleza, por ejemplo, algunos microorganismos se reproducen más a cierta temperatura, o a ciertos niveles de fosforo y nitrógeno. En la siguiente tabla se muestran cuáles son las características principales que debe tener un microorganismo para que sea considerado como bioindicador. (p. 139)

Tabla IV. **Características de un microorganismo bioindicador**

Característica
Fácil de aislar y cuantificar
No debe ser un patógeno
Existir en la flora intestinal de seres vivos sanos
Encontrarse en grandes cantidades
Resistencia a factores ambientales
Existir en las heces de los seres vivos
Fácil de aislar y cuantificar

Fuente: elaboración propia.

De entre los microorganismos más utilizados destacan los siguientes:

- Bacterias heterotróficas
- Coliformes fecales
- Coliformes totales

Para evaluar la calidad del agua en la presente investigación se utilizará la familia de coliformes fecales, las cuales destacan entre sus características: su comportamiento parecido a las bacterias que causan enfermedades, su

presencia tanto dentro como fuera de los seres vivos, y tienen una mayor capacidad de sobrevivencia que otras bacterias. (Diaz, Fall y Jimienez, 2003)

8. PROPUESTA DE ÍNDICE DE CONTENIDOS

ÍNDICE DE ILUSTRACIONES

LISTA DE SÍMBOLOS

GLOSARIO

RESUMEN

1. INTRODUCCIÓN
2. ANTECEDENTES
3. PLANTEAMIENTO DEL PROBLEMA
 - 3.1. Contexto general
 - 3.2. Descripción del problema
 - 3.3. Formulación del problema
 - 3.4. Delimitación del problema
4. JUSTIFICACIÓN
5. OBJETIVOS
 - 5.1. General
 - 5.2. Específicos
6. NECESIDADES POR CUBRIR Y ESQUEMA DE SOLUCIÓN
7. MARCO TEÓRICO
 - 7.1. Fundamentos estadísticos

- 7.2. Análisis de correlación
 - 7.2.1. Correlacion de Pearson
 - 7.2.2. Correlacion de Spearman
 - 7.2.3. Análisis de regresión
 - 7.2.3.1. Regresión lineal simple
 - 7.2.3.2. Regresión lineal múltiple
 - 7.2.3.3. Recta de regresión ajustada
 - 7.2.3.4. Comparación de modelos
 - 7.2.3.4.1. Coeficiente de determinación ajustado
 - 7.2.3.4.2. Estadístico CP de Mallows
- 7.3. Conceptos y definiciones
 - 7.3.1. Lago de Amatitlán
 - 7.3.1.1. Ríos de la cuenca
 - 7.3.2. Calidad del agua
 - 7.3.2.1. Microorganismos

8. PROPUESTA DE ÍNDICE DE CONTENIDOS

9. METODOLOGÍA

- 9.1. Características del estudio
- 9.2. Unidades de análisis
- 9.3. Variables e indicadores
- 9.4. Fases del estudio
 - 9.4.1. Fase 1: Revisión de la literatura
 - 9.4.2. Fase 2: Obtención y procesamiento de la información

- 9.4.3. Fase 3: Análisis de regresión
- 9.4.4. Fase 4: Análisis de secuencia temporal
- 9.4.5. Fase 5: Interpretación de los resultados
- 9.4.6. Fase 6: Elaboración de informe final

10. TÉCNICAS DE ANÁLISIS DE LA INFORMACIÓN

- 10.1. Técnicas metodológicas
 - 10.1.1. Revisión documental
 - 10.1.2. Meta-análisis
- 10.2. Técnicas estadísticas
 - 10.2.1. Medidas de tendencia central
 - 10.2.2. Medidas de dispersión
 - 10.2.3. Análisis gráfico
 - 10.2.4. Pruebas de hipótesis
 - 10.2.5. Análisis de residuos

11. CRONOGRAMA

12. FACTIBILIDAD DEL ESTUDIO

- 12.1. Recurso humano
- 12.2. Recursos financieros
- 12.3. Recursos tecnológicos
 - 12.3.1. *Software*
 - 12.3.2. *Hardware*
- 12.4. Acceso a la información

13. REFERENCIAS

14. APÉNDICE

15. ANEXO

9. METODOLOGÍA

9.1. Características del estudio

Para la presente investigación se utilizará un enfoque cuantitativo, ya que la variable de interés es la concentración de coliformes fecales, la cual es una variable numérica. De igual modo las variables independientes corresponden a parámetros numéricos ambientales, fisicoquímicos y químicos.

La investigación es de tipo exploratorio y descriptivo. Es exploratorio ya que anteriormente únicamente se había abordado el problema de la contaminación de los ríos de la cuenca del lago de Amatitlán con la estadística descriptiva. Es descriptivo, debido a que el estudio culminara con la presentación de las propiedades y características de los modelos estadísticos de regresión y series de tiempo construidos.

El alcance de la investigación es descriptivo, ya que el objetivo de la investigación es identificar los parámetros ambientales, fisicoquímicos y químicos de los ríos que más información aportan sobre la contaminación por coliformes fecales. También se busca describir el comportamiento que ha tenido la contaminación de los ríos en el periodo 2016 – 2020.

El presente estudio adoptará un diseño no experimental, debido a que la información ambiental, fisicoquímica y química de los ríos de la cuenca del lago de Amatitlán no tendrá ninguna clase de manipulación.

El estudio será transversal, ya que las observaciones sobre los ríos se tomarán como independientes entre sí cuando se estudie la relación de la contaminación con los parámetros de los ríos. El estudio también será longitudinal, ya que se estudiará la evolución de la contaminación de cada río en el tiempo, con el fin de realizar una proyección para los próximos años.

9.2. Unidades de análisis

La población de estudio serán las aguas de los principales ríos de la cuenca del lago de Amatitlán, específicamente los ríos Pampumay, Frutal, Pinula, Platanitos, San Lucas y Villalobos. Como es imposible físicamente analizar toda el agua de los ríos, se utilizará la información proveniente de las muestras mensuales tomadas por AMSA en el periodo 2016 a 2021. (AMSA, 2021)

9.3. Variables e indicadores

En todo estudio estadístico es necesario definir las variables dependientes e independientes que se utilizaran. En la siguiente tabla se puede observar que la única variable dependiente del estudio será la cantidad de coliformes fecales, la cual representa la cantidad de colonias de bacterias que hay en 100 mL. Por otro lado, se utilizarán 8 variables independientes, que corresponden a mediciones ambientales como la temperatura, mediciones químicas como la cantidad de oxígeno, fósforo y nitrógeno presente en el agua, y medidas fisicoquímicas como la conductividad del agua. Para cada variable se especifica la magnitud a medir, su dimensional y la escala de medición más adecuada.

Tabla V. Operativización de variables

Variables	Definición Teórica	Definición Operativa	Escala
Concentración de coliformes fecales (Y)	Medición de la cantidad de bacterias presentes en el agua	Valor numérico en UCF/100mL, representa las unidades formadoras de colonias por milímetros	De razón
Grado de acidez (X_1)	Medición del grado de acidez o alcalinidad de una sustancia	Valor numérico adimensional	De razón
Temperatura del agua (X_2)	Medida de la velocidad de promedio de las moléculas de agua	Valor numérico en °C, está basado en los puntos de fusión y ebullición del agua	De intervalo
Conductividad (X_3)	Capacidad de la materia para permitir el paso electrones	Valor numérico en uS/Cm, representa el inverso de un ohm	De razón
Oxígeno Disuelto (X_4)	Cantidad de oxígeno gaseoso que esta disuelto en el agua	Valor numérico en mg/L, representa cuántos miligramos de una solución están presentes en un litro de mezcla	De razón
Solidos totales disueltos (X_5)	Cantidad de todas las sustancias orgánicas e inorgánicas presentes en el agua en estado micro granular	Valor numérico en mg/L, representa el residuo que queda al evaporar un litro de agua filtrada	De razón
Porcentaje de salinidad (X_6)	Cantidad de sales minerales disueltas en un cuerpo de agua	Valor numérico porcentual (de 0 a 1),	De razón
Fósforo total (X_7)	Cantidad de todos los compuestos que contienen fósforo en el agua	Valor numérico en mg/L, representa cuantos miligramos de compuestos que contienen fósforo están presentes en un litro de agua	De razón
Nitrógeno total (X_7)	Cantidad de todos los compuestos que contienen nitrógeno en el agua	Valor numérico en mg/L, representa cuantos miligramos de compuestos que contienen nitrógeno están presentes en un litro de agua	De razón
Caudal (X_8)	Medida del volumen de agua que atraviesa una sección transversal en un tiempo determinado	Valor numérico en lts/seg, representa la cantidad de litros de agua que atraviesan el área de medición en un segundo	De razón

Fuente: elaboración propia.

9.4. Fases del estudio

El proceso para culminar exitosamente el estudio está compuesto por 4 fases, las cuales se describen a continuación:

9.4.1. Fase 1: Revisión de la literatura

Esta fase consiste en consultar fuentes bibliográficas con el fin de adquirir conocimientos sobre la historia, contexto y la naturaleza de la contaminación que sufren los ríos de interés. También se realizará una revisión de las técnicas estadísticas que se adecuen mejor al problema.

9.4.2. Fase 2: Obtención y procesamiento de la información

Esta fase consiste en obtener la información de la institución encargada del manejo sustentable de la cuenca (AMSA). Una vez obtenida la información necesaria para realizar la investigación, se limpiará la misma y se realizarán los cambios necesarios para homogenizar la estructura de la información.

9.4.3. Fase 3: Análisis de regresión

En esta fase se construirá un modelo preliminar de regresión multivariada utilizando todas las variables independientes, con dicho modelo se podrá identificar cuáles son las variables que más influyen en la variable de interés. Luego se construirá un modelo de regresión utilizando las variables independientes más influyentes y otro utilizando las variables independientes menos influyentes. Por último, se seleccionará el modelo de regresión más robusto utilizando los criterios de información AIC, BIC y el coeficiente de determinación ajustado.

9.4.4. Fase 4: Análisis de secuencia temporal

En esta fase se graficarán los datos de contaminación de cada río, dependiendo de las características de la secuencia de datos (tendencia, estacionalidad, estacionariedad) se construirá el modelo de suavizado exponencial más adecuado. También se construirá un modelo ARIMA de series de tiempo con los datos de cada río. Por último, se seleccionarán los modelos que realicen predicciones de los niveles de contaminación con la mayor robustez, para ello se utilizarán las métricas de error MSE, MAE y RMSE.

9.4.5. Fase 5: Interpretación de los resultados

En esta fase se analizarán e interpretarán los resultados obtenidos en el análisis de regresión y en el análisis temporal de la contaminación de los ríos.

9.4.6. Fase 6: Elaboración de informe final

Esta fase comprende la redacción y síntesis de todo el proceso metodológico utilizado para llevar a cabo la investigación, así como los resultados obtenidos, la interpretación y conclusiones de los mismos.

10. TÉCNICAS DE ANÁLISIS DE LA INFORMACIÓN

10.1. Técnicas metodológicas

Las técnicas metodológicas que se utilizarán para obtener la información necesaria para la investigación son:

10.1.1. Revisión documental

La información se obtendrá de datos históricos recopilados por la entidad encargada del manejo sustentable de la cuenca del lago de Amatitlán (AMSA) en estudios previos.

10.1.2. Meta-análisis

Se utilizará para comparar la compatibilidad entre la información proveniente de los distintos estudios realizados por AMSA, con el fin de resumir e integrar toda la información recopilada.

10.2. Técnicas estadísticas

Las técnicas estadísticas que se utilizarán para analizar y comprender la información en la investigación son:

10.2.1. Medidas de tendencia central

Se utilizarán las medidas de media, mediana y moda para describir el comportamiento de las variables independientes y dependientes.

10.2.2. Medidas de dispersión

Se utilizarán las medidas de máximo, mínimo, varianza y desviación estándar para describir la distribución de las variables independientes y dependiente.

10.2.3. Análisis gráfico

Se realizarán histogramas para conocer como es la distribución de las variables independientes y dependiente. Las gráficas de dispersión se utilizarán para visualizar la correlación entre las variables, también se usarán para evaluar la normalidad de las variables de forma visual. Por último, también se elaborarán correlogramas para evaluar la autocorrelación de los datos de contaminación.

10.2.4. Pruebas de hipótesis

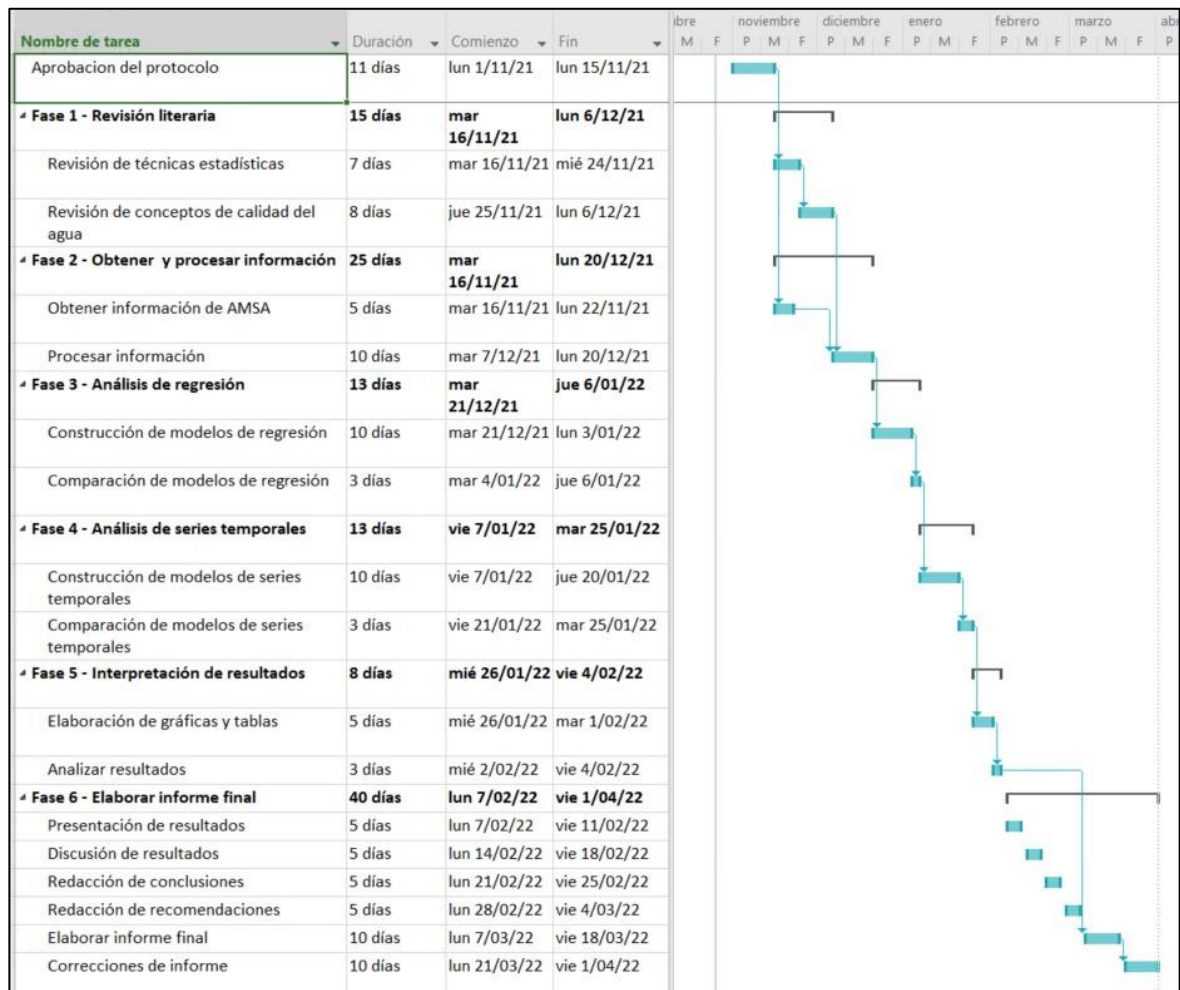
Se utilizarán pruebas de hipótesis para estimar la correlación entre las variables con un grado de confiabilidad. También se utilizarán para evaluar la normalidad de los datos.

10.2.5. Análisis de residuos

Para los modelos de regresión y series de tiempo se evaluará si los residuos generados cumplen los supuestos de normalidad, homocedasticidad e independencia.

11. CRONOGRAMA

Figura 7. Cronograma de actividades



Fuente: elaboración propia, realizado con Microsoft Project.

12. FACTIBILIDAD DEL ESTUDIO

12.1. Recurso humano

El recurso humano necesario para la culminación y presentación exitosa de la investigación está conformado por:

- Estudiante investigador
- Asesor
- Revisor de lingüística externo

12.2. Recursos financieros

Una parte muy importante del planteamiento de una investigación es la viabilidad de la misma. Por muy buena que sea la idea base de un estudio, si el costo de realizarlo es muy elevado, el estudio es muy poco probable que se lleve a cabo. En la siguiente tabla se muestra el detalle del costo de los materiales y herramientas necesarias para concluir exitosamente la investigación.

Tabla VI. Recursos financieros

Elemento	Unidad	Costo unitario (Q)	Cantidad necesaria	Costo (Q)
Servicio de internet	Mes	259.00	4	1,036.00
Consumo de energía eléctrica	Mes	200.00	4	800.00
Licencia de paquete Office	Licencia anual	469.99	1	469.99
Papel carta	Resma	34.90	2	69.84
Software estadístico	Programa	0.00	1	0.00
Licencia de Lucidchart	Licencia anual	667.80	1	667.80
Impresora	Equipo	219.00	1	219.00
Tinta	Cartucho	65.00	2	130.00
Computadora personal	Equipo	0.00	1	0.00
Servicio de revisor externo	Servicio	500	1	500.00
Teléfono	Equipo	0.00	1	0.00
Plan telefónico	Mes	150.00	2	300.00
			Total	4,192.63

Fuente: elaboración propia.

La investigación será financiada completamente por el estudiante.

12.3. Recursos tecnológicos

Los recursos tecnológicos necesarios para finalizar la investigación se clasifican de la siguiente manera:

12.3.1. Software

R Studio, Sistema Operativo Windows 10, paquete Microsoft Office, Google Chrome, Infostat, software de diagramación Lucidchart, servicio de correo y almacenamiento en la nube.

12.3.2. Hardware

Computadora personal, teléfono celular, impresora.

12.4. Acceso a la información

La información es de acceso público y se obtendrá de la entidad encargada del manejo sustentable de la cuenca del lago de Amatitlán (AMSA).

13. REFERENCIAS

1. Aparicio, J., Martínez, A., y Morales, J. (2004). *Modelos lineales aplicados en R. Elche: Centro de Investigación Operativa.*
2. Bautista Olivas, A. L., Tovars Salinas, J. L., Mancilla Villa, Ó. R., Magdaleno Flores, H. E., Ramírez Ayala, C., y Arteaga Ramírez, R. (2013). *Calidad microbiológica del agua obtenida por condensación de la atmosfera en Tlaxcala, Hidalgo y Ciudad de México.*
3. Cano Alfaro, M. F. (2018). *Diagnóstico de los recursos hídricos de la cuenca del lago de Amatitlán.* Amatitlán: AMSA.
4. Caycho, C., Castillo, C., y Merino, V. (2020). *Manual de estadística no paramétrica aplicada a los negocios.*
5. Costa, M. A. (2015). *Statistical modelling of water quality time series – the river Vouga Basin case study. Research and Practices in Water.* Recuperado de <https://doi.org/10.5772/59062>.
6. Decreto número 64-96. Ley de creación de la autoridad para el manejo sustentable de la cuenca y del lago de Amatitlán. Congreso de la República de Guatemala. 18 de septiembre de 1996.

7. Díaz Delgado, C., Fall, C., y Jiménez, Q. (2003). *Agua potable para comunidades rurales, reusó y tratamientos avanzados de aguas residuales domésticas*. Ciudad de México: Iberoamericana de Potabilización y Depuración del Agua.
8. Dirección General de Obras Hidráulicas y Calidad de las Aguas. (2000). *Libro blanco del agua en España*. Madrid: Centro de publicaciones Ministerio de Ambiente.
9. División de Control y Calidad Ambiental. (2020). *Boletín técnico: Calidad del agua en el lago de Amatitlán*.
10. Fernández Santiesteban, M. T. (2017). *Determinación de coliformes totales y fecales en aguas de uso tecnológico para las centrifugas. ICIDA. Sobre los Derivados de la Caña de Azúcar*.
11. Gamboa Becerra, R. A., Cifuentes Osorio, G. R., y Rocha Gil, Z. E. (2016). *Indicadores bacterianos de contaminación fecal en el agua del embalse La Copa, municipio de Toca, Boyacá/Colombia*. 13+, 10-23. Recuperado de <https://doi.org/10.24267/23462329.157>.
12. Morantes Quintana, R., Polo, G., y Pérez Santodomingo, N. (2019). *Modelo de regresión lineal múltiple para estimar la concentración de PM1. Revista Internacional de Contaminación Ambiental*.
13. Navidi, W. (2006). *Estadística para ingenieros*. México: McGraw-Hill.
14. Novales, A. (2010). *Análisis de regresión*. Madrid: Universidad Complutense.

15. Pérez Gudiel, D. B. (2007). *Evaluación del efecto de la aireación artificial para mejorar la calidad del agua del lago de Amatitlán*. Guatemala: Universidad de San Carlos de Guatemala.
16. Pérez, J. I., Nardini, A. G., y Galindo, A. A. (2018). *Análisis comparativo de índices de calidad del agua aplicados al río Ranchería, La Guajira-Colombia*. Recuperado de <https://dx.doi.org/10.4067/S0718-07642018000300047>.
17. Pisarra, T. C., Valera, C. A., Costa, R. C., Siqueira, H. E., Martins Filho, M. V., y Pacheco, F. A. (2019). *A regression model of stream water quality base on interactions between landscape composition and riparian buffer width in small catchments*. *Water*. Recuperado de <https://doi.org/10.3390/w11091757>.
18. Soo, Y., y Seo, W. (2018). *Prediction of fecal coliform using logistic regression and tree-based classification models in the North Han River, South Korea*. *Journal of Hydro-environment Research*. Recuperado de <https://doi.org/10.1016/j.jher.2018.09.002>.
19. Taheri, T., Ghashaghaie, M., y Georgiou, P. (2014). *Time series analysis of water quality parameters*. *Journal of Applied Research in Water and Wastewater*, 43-52.
20. Vega Araya, M., y Alvarado Barrantes, R. (2019). *Análisis de series de tiempo de variables biofísicas para cuatro regiones de Guanacaste, Costa Rica*. *Revista de Ciencias Ambientales*, 60-96.

21. Walpole, R. E., Myers, R. H., Myers, S. L., y Ye, K. (2012). *Probabilidad y estadística para ingeniería y ciencias*. México: Pearson Education.

22. Zimmer Faust, A. G., Brown, C. A., y Manderson, A. (2018). *Statistical models of fecal coliform levels in Pacific Northwest estuaries for improved shellfish harvest area closure decision making*. *Marine Pollution Bulletin*, 360-369. Recuperado de <https://doi.org/10.1016/j.marpolbul.2018.09.028>.

14. APÉNDICE

Apéndice 1. Matriz de coherencia

No.	Preguntas de investigación	Objetivos	Metodología
1	¿Cuál es el grado de correlación que existe entre el nivel de coliformes fecales de los ríos de la cuenca con sus factores ambientales, fisicoquímicos y químicos?	Estimar el nivel de correlación entre la concentración de coliformes fecales y los factores ambientales, fisicoquímicos y químicos a través del coeficiente de Pearson, para identificar el tipo de relación de cada uno de los factores.	Se calculará el coeficiente de correlación de Pearson entre cada uno de los factores con la concentración de coliformes fecales presentes en el agua.
2	¿Qué variables ambientales, fisicoquímicas y químicas aportan más información sobre la concentración de coliformes fecales en los ríos de la cuenca?	Identificar las variables ambientales, fisicoquímicas y químicas que más información aporta sobre la concentración de coliformes fecales mediante un análisis de regresión, para descartar las variables menos influyentes.	Se realizará un análisis de regresión entre los factores de los ríos con la concentración de coliformes fecales. Luego por medio de los valores CP de Mallows, el intervalo y la magnitud de los coeficientes de cada variable se identificará cuáles son las variables que aportan menor información al modelo de regresión.
3	¿Cuál es la robustez de los modelos de regresión que relacionan la concentración de coliformes fecales con los parámetros ambientales, fisicoquímicos y químicos de los ríos de la cuenca?	Calcular la robustez de los modelos de regresión que relacionan la concentración de coliformes fecales con los parámetros ambientales, fisicoquímicos y químicos mediante los indicadores AIC, BIC y R^2 ajustado. Para identificar cual es el modelo que mejor se ajusta a los datos.	Se construirán varios modelos de regresión, empleando cada uno un conjunto distinto de variables independientes, para comparar cuál de todos los modelos es el que mejor se ajusta a los datos se utilizaran los criterios AIC, BIC y el coeficiente de determinación ajustado.
4	¿Cuál será el nivel de coliformes fecales de los principales ríos de la cuenca del lago de Amatitlán en los años 2022 y 2023?	Estimar el nivel de coliformes fecales de los ríos de la cuenca para los años 2022 – 2023 mediante la construcción de un modelo ARIMA, para conocer cuál será el estado de los ríos en los próximos años.	Se utilizarán los datos del nivel de coliformes fecales presentes en el agua de los ríos en el periodo 2016 – 2021 para construir un modelo ARIMA por cada río. Los modelos se construirán minimizando el valor de la raíz cuadrada del error cuadrático medio (RMSE).

Fuente: elaboración propia.

15. ANEXO

Anexo 1. Análisis de anti plagio

2340463090114_2.pdf (27/10/20) x +

plagscan.com/doc?142416147&sharekey=rNBRkiTkYLWHshMvLCVK

PlagScan by Ouriginal. Resultados del Análisis de los plagios del 2340463090114_2.pdf

27/10/2021 21:34 Fecha: 27/10/2021 21:27

1.6%

Vista: * Todas las fuentes 4

11 resultados

Todas las fuentes 4

Top tres 3

Fuentes de internet 4

- [0] ichi.pro/es/comprencion-de-la-regresioi 0.7% 5 resultados | Marcar resultados en la t
- [1] www.researchgate.net/publication/2812 0.6% 3 resultados | Marcar resultados en la t
- [2] library.co/document/qvvg4x0q-selecci 0.2% 2 resultados | Marcar resultados en la t
- [3] books.google.co.in/books?id=jVhBAU 0.4% 1 resultados | Marcar resultados en la t

Leyenda marcado del texto

- Aa concordancia exacta
- Aa cambios del texto posibles
- Aa marcado como cita

1. ANTECEDENTES

La contaminación de un cuerpo de agua es un tema que ha tenido importancia desde hace décadas, desde que se comenzó a estudiar el crecimiento poblacional y el aumento de la urbanización se empezó a cuestionar el impacto que dichos fenómenos sociales generarían en el medio ambiente.

Entre los estudios recientes de la calidad del agua en la cuenca del lago podemos mencionar el realizado por Cano (2018); en su informe se realiza un análisis descriptivo

Fuente: Plagscan. *Análisis de anti plagio*. Consultado el 27 de octubre de 2021. Recuperado de <https://www.plagscan.com/doc?142416147&sharekey=rNBRkiTkYLWHshMvLCVK>.

