Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Artes en Estadística Aplicada

PROPUESTA DE UN MODELO ESTADÍSTICO GENERADO A TRAVÉS DE ANÁLISIS MULTIVARIADO PARA ESTIMAR LA CALIDAD DEL AGUA DE LOS RÍOS XAYÁ Y PIXCAYÁ DE GUATEMALA

Lic. Bryant Barrientos Castellanos
Asesorado por la Maestra Norma Edith Gil Rodas

Guatemala, julio de 2024

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

PROPUESTA DE UN MODELO ESTADÍSTICO GENERADO A TRAVÉS DE ANÁLISIS MULTIVARIADO PARA ESTIMAR LA CALIDAD DEL AGUA DE LOS RÍOS XAYÁ Y PIXCAYÁ DE GUATEMALA

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

LIC. BRYANT BARRIENTOS CASTELLANOS
ASESORADO POR LA MAESTRA NORMA EDITH GIL RODAS

AL CONFERÍRSELE EL TÍTULO DE

MAESTRO EN ARTES EN ESTADÍSTICA APLICADA

GUATEMALA, JULIO DE 2024

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. José Francisco Gómez Rivera (a. i.)
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Ing. Kevin Vladimir Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. José Francisco Gómez Rivera (a. i.)
EXAMINADORA	Mtra. Inga. Aurelia Anabella Cordova Estrada
EXAMINADOR	Mtro. Ing. Edwin Adalberto Bracamonte Orozco
EXAMINADORA	Dra. Aura Marina Rodríguez Pérez
SECRETARIO	Mtro. Ing. Hugo Humberto Rivera Pérez

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:

PROPUESTA DE UN MODELO ESTADÍSTICO GENERADO A TRAVÉS DE ANÁLISIS MULTIVARIADO PARA ESTIMAR LA CALIDAD DEL AGUA DE LOS RÍOS XAYÁ Y PIXCAYÁ DE GUATEMALA

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Postgrado, con fecha 31 de enero de 2022.

Lic. Bryant Barrientos Castellanos

USAC
TRICENTENARIA
Universidad de San Carlos de Guatemala

Decanato
Facultad e Ingeniería

24189101- 24189102

LNG.DECANATO.OIE.375.2024

El Decano 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 Estudios de Posgrado, al Trabajo de Graduación titulado: **PROPUESTA DE UN MODELO ESTADÍSTICO GENERADO A TRAVÉS DE ANÁLISIS MULTIVARIADO PARA ESTIMAR LA CALIDAD DEL AGUA DE LOS RÍOS XAYÁ Y PIXCAYÁ DE GUATEMALA**, presentado por: **Bryant Barrientos Castellanos** después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

Firmado electrónicamente por: José Francisco Gómez Rivera
Motivo: Autorización de entrega de trabajo de graduación
Fecha: 12/08/2024 19:14:39
Lugar: Facultad de Ingeniería, USAC.

Ing. José Francisco Gómez Rivera
Decano a.i.

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANO a.i.
Facultad de Ingeniería

Guatemala, julio de 2024

Para verificar validez de documento ingrese a https://www.ingenieria.usac.edu.gt/firma-electronica/consultar-documento

Tipo de documento: Correlativo para orden de impresión Año: 2024 Correlativo: 375 CUI: 2545756680101

Escuelas: Ingeniería Civil, Ingeniería Mecánica Industrial, Ingeniería Química, Ingeniería Mecánica Eléctrica, - Escuela de Ciencias, Regional de Ingeniería Sanitaria y Recursos Hidráulicos (ERIS). Postgrado Maestría en Sistemas Mención Ingeniería Vial. Carreras: Ingeniería Mecánica, Ingeniería Electrónica, Ingeniería en Ciencias y Sistemas. Licenciatura en Matemática. Licenciatura en Física. Centro de Estudios Superiores de Energía y Minas (CESEM). Guatemala, Ciudad

Guatemala, julio de 2024

LNG.EEP.OID.2.2024

En mi calidad de DIRECTORA de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

"PROPUESTA DE UN MODELO ESTADÍSTICO GENERADO A TRAVÉS DE ANÁLISIS MULTIVARIADO PARA ESTIMAR LA CALIDAD DEL AGUA DE LOS RÍOS XAYÁ Y PIXCAYÁ DE GUATEMALA"

presentado por **BRYANT BARRIENTOS CASTELLANOS** correspondiente al programa de **MAESTRÍA EN ESTADÍSTICA APLICADA** ; apruebo y autorizo el mismo.

Atentamente,

"ID Y ENSEÑAD A TODOS"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DIRECTORA
POSTGRADO

MAESTRA Aurelia Anabela Cordova Estrada
DIRECTORA
Escuela de Estudios de Postgrado

Guatemala, 26 de octubre de 2023

M.A. Inga. Aurelia Anabela Cordova Estrada
Directora
Escuela de Estudios de Postgrado
Presente

Estimada M.A. Inga. Cordova Estrada

Por este medio informo a usted, que he revisado y aprobado el **INFORME FINAL y ARTÍCULO CIENTÍFICO** titulado: **PROPUESTA DE UN MODELO ESTADÍSTICO GENERADO A TRAVÉS DE ANÁLISIS MULTIVARIADO PARA ESTIMAR LA CALIDAD DEL AGUA DE LOS RÍOS XAYÁ Y PIXCAYÁ DE GUATEMALA** del estudiante **Bryant Barrientos Castellanos** quien se identifica con número de carné **200810230** del programa de Maestria En Estadistica Aplicada.

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el ***Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014.*** Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

UNIVERSIDAD DE SAN CARLOS
ESCUELA DE POSTGRADO
FACULTAD DE INGENIERIA
DE GUATEMALA

Msc. Ing. Edwin Adalberto Bracamonte Orozco
Coordinador
Maestria En Estadistica Aplicada
Escuela de Estudios de Postgrado

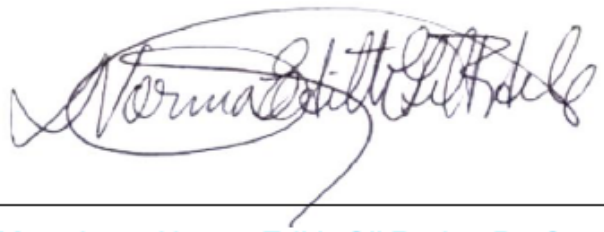
Guatemala, 26 de octubre de 2023

M.A. Inga. Aurelia Anabela Cordova Estrada
Directora
Escuela de Estudios de Postgrados
Presente

Estimada M.A. Inga. Cordova Estrada

Por este medio informo a usted, que he revisado y aprobado el Trabajo de Graduación y el Artículo Científico: **"PROPUESTA DE UN MODELO ESTADÍSTICO GENERADO A TRAVÉS DE ANÁLISIS MULTIVARIADO PARA ESTIMAR LA CALIDAD DEL AGUA DE LOS RÍOS XAYÁ Y PIXCAYÁ DE GUATEMALA"** de el/la estudiante **Bryant Barrientos Castellanos** del programa de **Maestría En Estadística Aplicada** identificado(a) con número de carné 200810230.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.



Mtra. Inga. Norma Edith Gil Rodas De Castillo
Colegiado No. 1375
Asesora de Tesis

ACTO QUE DEDICO A:

Mis padres	Gilberto Barrientos (q.e.p.d.) y Patricia Castellanos, por cuyo incansable esfuerzo y dedicación, que únicamente puede ser explicado por el amor más puro, he llegado a donde estoy; no sería nadie sin ustedes.
Mis hermanos	Kristhell, Jeffrey y Kevin Barrientos, quienes a través de su amor y apoyo han motivado mi superación en todo momento.
Mi familia	Por ser un ejemplo de amor, unidad y superación.
Mis amigos	Pablo, Francisco y Carol, por ser cómplices en las muchas locuras que hemos hecho en esta vida, pues, ¿qué sería de esta vida sin un poco de diversión?
Mis mentores	Por haberme permitido ver la belleza de la ciencia y ponerla al servicio de la sociedad. Especialmente a Bessie Oliva y a Norma Gil, quienes sin saber construyeron en mí la pasión por la química ambiental.
Soluciones Analíticas	Por ser uno de los pilares más importantes en mi formación profesional.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por la inconmensurable oportunidad de acceso a la formación académica profesional.
Pueblo de Guatemala	Por quien el proyecto de educación superior de la Universidad de San Carlos de Guatemala es posible; del pueblo es y al pueblo se debe.
Mi familia	Por la paciencia, los ánimos y sobre todo la confianza. El “más vale tarde que nunca” ha llegado a su fin y el ciclo se cierra gracias a ustedes.
Norma Gil	Por todo lo entregado como asesora en este proyecto. Sin su tiempo, aportes y apoyo esto no hubiera sido posible.
Gabriela Castellanos y Axel Juárez	Por todos los recordatorios y llamados a culminar este proyecto, por aquellos mensajes positivos y de confianza y por recordarme a menudo que el poder está en uno.
Luis Orellana	Por haber confiado en mí y darme la oportunidad de crecer profesionalmente.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES V
LISTA DE SÍMBOLOS VII
GLOSARIO IX
RESUMEN XIII
PLANTEAMIENTO DEL PROBLEMA XV
OBJETIVOS XIX
RESUMEN DEL MARCO METODOLÓGICO XXI
INTRODUCCIÓN XXVII

1. MARCO REFERENCIAL 1

2. MARCO TEÓRICO 5
2.1. Fundamentos matemáticos y estadísticos 5
2.1.1. Álgebra lineal 5
2.1.1.1. Vectores 6
2.1.1.2. Matrices 6
2.1.1.3. Operaciones de vectores y matrices 7
2.1.2. Conceptos estadísticos básicos 11
2.1.2.1. Medidas de localización 11
2.1.2.2. Medidas de variabilidad 13
2.1.2.3. Distribuciones continuas de probabilidad 15
2.1.3. Regresión lineal 18
2.1.3.1. Curvas de aproximación 19
2.1.3.2. Modelo de regresión lineal 20

2.1.3.3. Método de mínimos cuadrados 22
2.1.3.4. Inferencias sobre los coeficientes..... 23
2.1.3.5. Coeficientes de correlación y determinación 24
2.1.4. Análisis estadístico multivariado..... 26
2.1.4.1. Matrices de datos 26
2.1.4.2. Análisis de componentes principales..... 27
2.1.4.3. Análisis factorial..... 33
2.2. Índices de calidad o contaminación del agua..... 35
2.2.1. Definición..... 35
2.2.2. Características. 36
2.2.3. Tipos de índices 37
2.2.3.1. Por el significado de la escala 37
2.2.3.2. Por la metodología utilizada para su creación..... 38
2.2.3.3. Por sus componentes..... 39
2.2.3.4. Por su regla de agregación 39
2.2.4. Creación..... 40
2.2.4.1. Selección de parámetros..... 40
2.2.4.2. Transformación de parámetros..... 42
2.2.4.3. Asignación de pesos 43
2.2.4.4. Agregación de subíndices 45
2.3. Marco legal..... 47
2.3.1. Constitución Política de la República de Guatemala..... 48
2.3.2. Ley de Protección y Mejoramiento del Medio Ambiente 48
2.3.3. Código de Salud..... 49
2.3.4. Agua para consumo humano 49

2.3.5. Agua envasada para consumo humano 50
2.3.6. Acuerdo Gubernativo 137-2006..... 50
2.3.7. Acuerdo Gubernativo 236-2006..... 51
2.3.8. Otra reglamentación 51

3. RESULTADOS..... 53
3.1. Objetivo 1. Seleccionar las variables influyentes mediante la reducción de la dimensionalidad, con las técnicas de análisis de componentes principales y análisis factorial. 53
3.1.1. Exploración de la base de datos..... 53
3.1.1.1. Base de datos sin tratamiento 53
3.1.1.2. Base de datos estandarizada 54
3.1.1.3. Estadística descriptiva 56
3.1.2. Estandarización de escala..... 58
3.1.3. Análisis de componentes principales..... 60
3.1.3.1. Correlaciones entre variables 60
3.1.3.2. Multicolinealidad 61
3.1.3.3. Adecuación del muestreo 62
3.1.3.4. Independencia de las variables 62
3.1.3.5. Componentes principales 62
3.2. Objetivo 2. Asignar pesos a las variables influyentes por medio de la aplicación de análisis factorial a los datos de las variables influyentes preseleccionadas con el análisis de componentes principales. 65
3.2.1. Valores propios..... 65
3.2.2. Componentes rotados 66
3.2.3. Pesos temporales..... 67
3.3. Objetivo 3. Transformar la escala de las variables seleccionadas, tomando como base los distintos

"niveles" deseables que indica la literatura para las variables de estudio.....68
3.4. Objetivo 4. Establecer la estructura del modelo y evaluar su confiabilidad mediante la comparación con otros índices de calidad aceptados mundialmente y criterios técnicos.....69
3.4.1. Estructura matemática y su comparación.....70
3.4.2. Estructura categórica y su comparación 71
3.5. Objetivo general. Proponer un modelo para estimar la calidad fisicoquímica del agua de los ríos Xayá y Pixcayá en Guatemala por medio de estadística multivariada. 72

4. DISCUSIÓN DE RESULTADOS.....75
4.1. Análisis interno.....75
4.2. Validez externa 88

CONCLUSIONES.....93
RECOMENDACIONES 95
REFERENCIAS.....97
APÉNDICES 103

ÍNDICE DE ILUSTRACIONES

FIGURAS

Figura 1. Gráficos de caja de las variables con valores atípicos 56
Figura 2. Gráficos de caja de las variables sin valores atípicos 57
Figura 3. Histogramas de las variables estandarizadas 59
Figura 4. Gráficos de caja de las variables estandarizadas 60
Figura 5. Correlograma con correlación de Spearman 61
Figura 6. Valoración y categorización de la calidad 70

TABLAS

Tabla 1. Operativización de variables XXII
Tabla 2. Funciones matemáticas para curvas de aproximación 20
Tabla 3. Resumen de variables y valores faltantes 55
Tabla 4. Estadísticos descriptivos 58
Tabla 5. Valores propios y varianza explicada 63
Tabla 6. Puntajes de los componentes principales 64
Tabla 7. Valores propios y varianza explicada ACP - Influyentes 65
Tabla 8. Análisis de factores por rotación Varimax 66
Tabla 9. Varianza explicada por los componentes rotados 66
Tabla 10. Puntajes al cuadrado 67
Tabla 11. Puntajes escalados a la unidad 67
Tabla 12. Pesos relativos 68
Tabla 13. Valores máximos de tolerancia 69
Tabla 14. Categorías de calidad del agua 71

LISTA DE SÍMBOLOS

Símbolo	Significado
Ca	Calcio
Cl	Cloruro
a_i	Coeficiente de ajuste
r	Coeficiente de correlación producto momento de Pearson
r^2	Coeficiente de determinación
PC_i	Componente principal
a_i^2	Comunalidad
s	Desviación estándar muestral
s_f	Desviación media absoluta
F	Factor común
$°C$	Grados Celsius
Fe	Hierro
ICA_e	Índice específico de calidad del agua
ICA_g	Índice global de calidad del agua
Mg	Magnesio
$\overline{x}$	Media
m_f	Media de la variable transformada logarítmicamente
$\frac{\mu S}{cm}$	Microsiemens por centímetro
$\frac{mg}{L}$	Miligramo por litro
NO_3^-	Nitrato

NO_2^-	Nitrito
$\frac{NMP}{mL}$	Número más probable por mililitro
w_i	Peso del subíndice
q_i	Subíndice o variable aleatoria transformada
SO_4^{2-}	Sulfato
$ud.pH$	Unidades de pH
$ud.Pt-Co$	Unidades de platino-cobalto
UNT	Unidades nefelométricas de turbidez
λ_i	Valor propio
P_m	Variabilidad acumulada
x_i	Variable aleatoria (también parámetro)
z_{if}	Variable aleatoria que ha sido transformada logarítmicamente y luego estandarizada
s^2	Varianza muestral
γ_i	Vector propio

GLOSARIO

Abióticos	Referente a los objetos sin vida
AC	Análisis de conglomerados o análisis de agrupamiento.
ACP	Análisis de componentes principales
AEM	Análisis estadístico multivariado
AF	Análisis factorial
Agentes biológicos	Organismos bióticos, en este documento organismos superiores como plantas, animales y otros que se pueden ver a simpe vista.
Agentes físicos	Factores abióticos que no se incluyen dentro de los agentes químicos.
Agentes químicos	Factores abióticos característicos por tener una estructura química definida y conocida.
Biológica(o)	Referente a los agentes biológicos
Biótico	Referente a organismos vivos

Cianobacteria	Microorganismo capaz de realizar fotosíntesis
CCPP	Componentes principales
CCRR	Componentes rotados
CE	Conductividad eléctrica
COGUANOR	Comisión Guatemalteca de Normas
COL	Color
CP	Componente principal
CR	Componente rotado
Contaminación	Introducción de agentes físicos, químicos o microbiológicos al ambiente capaces de alterarlo de manera nociva.
Cuenca	Conjunto de cuerpos de agua que afluyen a un lugar común.
Cuerpo léntico	Cuerpo de agua que no fluye
Cuerpo lótico	Cuerpo de agua que fluye
DUR	Dureza

Eutrofización	Incremento de nutrientes, principalmente fósforo y nitrógeno, en las aguas dulces que provoca un aumento de fitoplancton.
FAO	Organización de las Naciones Unidas para la Alimentación y la Agricultura.
Fisicoquímica(o)	Referente a los agentes físicos o químicos
Fitoplancton	Algas microscópicas
INE	Instituto Nacional de Estadística
KMO	Prueba de Kaiser-Meyer-Olkin
LMA	Límite máximo admisible, valor por arriba del cual se perciben cambios sensoriales en el agua, pero que no afectan la salud.
LMP	Límite máximo permisible, valor por arriba del cual, además de poderse percibir cambios sensoriales, existe un riesgo a la salud.
Macroinvertebrados	Insectos, moluscos y otros invertebrados que habitan en lugares con agua dulce.
Microbiológica(o)	Referente a los agentes microbiológicos

NSF-WQI	Índice de calidad del agua de la National Sanitation Foundation
OMS	Organización Mundial de la Salud
pH	Potencial de hidronio
Polución	Forma intensa de contaminación
SDT	Sólidos disueltos totales
TEM	Temperatura
TUR	Turbidez

RESUMEN

El propósito de este estudio fue la creación de un índice de calidad del agua para las cuencas Xayá – Pixcayá, el cual funge como una herramienta capaz de poner al alcance de la comprensión de toda la sociedad la información científica relacionada, consolidándola en un número similar a una calificación de examen. Promoviendo así la concienciación social con respecto a la problemática de contaminación en las cuencas.

El enfoque adoptado para ello fue cuantitativo, con diseño no experimental de tipo descriptivo correlacional y alcance transversal.

Se generó un modelo para estimar la calidad del agua en las cuencas Xayá – Pixcayá, denominado Índice de Calidad del Agua Xayá – Pixcayá 2023 (o ICAXP23 de forma abreviada), mediante el auxilio de técnicas estadísticas multivariantes.

La comparación con otros índices permitió concluir que el desempeño del índice propuesto fue adecuado. Se recomienda mejorar el plan de monitoreo en la cuenca, tanto en frecuencia como en la selección de variables, para poder aumentar el tamaño de muestra y mejorar el desempeño de las técnicas de análisis de componentes principales y análisis factorial.

PLANTEAMIENTO DEL PROBLEMA

Contexto general

Guatemala tiene diversos programas de monitoreo de parámetros de calidad del agua en cuerpos lóticos, llevados a cabo por instancias públicas, académicas y privadas con poca o nula divulgación social. Entre estos programas de monitoreo se encuentran aquellos que realizan el Instituto Nacional de Sismología, Vulcanología, Meteorología e Hidrología (INSIVUMEH), la Autoridad para el Manejo Sustentable de la Cuenca y el Lago de Amatitlán (AMSA), la Autoridad para el Manejo Sustentable de la Cuenca del Lago de Atitlán y su Entorno (AMSCLAE), las diversas investigaciones de las universidades públicas y privadas del país, entre otros.

La calidad de agua de los ríos de Guatemala es en general baja, ya sea por situaciones naturales o antropogénicas. Los ríos, que son las principales fuentes de agua para uso y consumo humano, recorren grandes distancias y tienen el potencial de trasladar contaminantes de un área geográfica a otra. Guatemala es el país con mayor cantidad de recursos hídricos de Centroamérica y a la fecha cuenta con problemas coyunturales internos y demandas sociales locales e internacionales en torno a la contaminación del recurso hídrico.

En materia legal, Guatemala cuenta con la Ley de Protección y Mejoramiento del Medio Ambiente y el Reglamento de Descargas y Reúso de Aguas Residuales y de la Disposición de Lodos. Sin embargo, ha habido múltiples prórrogas al cumplimiento del reglamento, especialmente para las

municipalidades, y no parece haber voluntad de las autoridades para su cumplimiento.

Guatemala es un país en el que la educación secundaria y universitaria es un privilegio al alcance de pocos. Además, en el sector público el sistema educativo de nivel primario y secundario está en un deterioro evidente al que se suma la dificultad de acceso por parte de la población que se encuentra en la línea de pobreza o pobreza extrema. Así, la información científica es un recurso al alcance comprensivo únicamente de académicos; lo que provoca que la concientización sobre el manejo de residuos y su disposición final sea un hito por cumplir de la sociedad con implicaciones en la calidad de los recursos hídricos del país.

Durante el 2021, el Comité Técnico de Normalización de Ambiente para la Calidad del Agua, de la Comisión Guatemalteca de Normas (COGUANOR) trabaja en la adopción, modificación o creación de normas en pro del recurso hídrico del país, con énfasis al uso sustentable de este y al mantenimiento de las cuencas hídricas. Dentro de las líneas de trabajo del Comité se encuentra el Proyecto de Norma NTG 77003 *Calidad del Agua. Evaluación y clasificación de la calidad de cuerpos superficiales. Índice de calidad de agua para ríos*. El Comité crea el índice de calidad de agua a través del método Delphi, un método basado en la lógica difusa que, a través de una serie de iteraciones de la opinión de un panel de expertos busca encontrar “la verdad” con respecto a un tema en específico.

Descripción del problema

Los recursos hídricos del país se encuentran deteriorados y siguen sufriendo de deterioro. El clamor de diversos sectores de la sociedad y de otros

países se han hecho evidentes y constituyen un problema nacional en materia coyuntural y de demandas internacionales.

El desconocimiento generalizado de las implicaciones económicas, sociales y en salud pública que acarrea la contaminación de los cuerpos de agua aunado a una palpable falta de voluntad política de cumplir con las normativas o leyes en materia ambiental ya existentes y a la falta de generación de nuevas normativas ambientales provoca que la contaminación en los cuerpos de agua de Guatemala aumente.

Además, la labor de informar y generar conciencia resulta ardua en una sociedad con escaso acceso a educación superior y, aún en personas con acceso a educación superior, el lenguaje técnico de la biología, la microbiología y la química, así como los resultados que de estas disciplinas se derivan resultan difíciles de interpretar para las personas que no tienen una preparación académica en estas disciplinas o en disciplinas afines.

Guatemala no cuenta con ninguna herramienta que permita divulgar información científica concerniente a la calidad del agua de los distintos cuerpos de agua del país a toda la población; lo que limita su acceso únicamente a los académicos.

Una opción viable para solventar esta situación es la construcción de un índice de calidad de agua. Los índices pueden ser construidos de diversas maneras, sin embargo, la construcción probabilística es especialmente interesante porque permite suplir las deficiencias inherentes, tanto a los índices, como a las técnicas de construcción por técnicas iterativas o de lógica difusa.

Un índice de calidad del agua se puede construir a través del análisis multivariado de los datos de análisis químicos, físicos o microbiológicos. Guatemala no cuenta con ningún índice de calidad del agua para ríos.

Formulación del problema

Pregunta central

¿Es posible proponer un modelo que permita estimar el grado de calidad del agua en los ríos Xayá y Pixcayá en Guatemala con base en los distintos parámetros de calidad del agua (variables regresoras)?

Preguntas auxiliares

- ¿Cuáles son las variables (parámetros) más influyentes dentro del modelo para la determinación de un índice de calidad fisicoquímico del agua?

- ¿Cuál es la importancia relativa de cada variable?

- ¿Qué anamorfismos de las variables (parámetros) permiten una transformación de escala que se alineen con las bases científicas de la química y la microbiología?

- ¿Cuál es la estructura del modelo que permite estimar la calidad fisicoquímica del agua en los ríos Xayá y Pixcayá de Guatemala a través de estadística?

Delimitación del problema

El objeto de estudio para este problema son los indicadores de calidad del agua de los ríos Xayá y Pixcayá de Guatemala, de los años 2007 a 2021, para los cuales se cuenta con información de monitoreo rutinaria.

OBJETIVOS

General

Proponer un modelo generado a través de análisis multivariado para estimar la calidad del agua de los ríos Xayá y Pixcayá de Guatemala.

Específicos

1. Seleccionar las variables influyentes mediante la reducción de la dimensionalidad, con las técnicas de análisis de componentes principales.
2. Asignar pesos a las variables influyentes por medio de la aplicación de análisis factorial a los datos de las variables influyentes preseleccionadas con el análisis de componentes principales.
3. Transformar la escala de las variables seleccionadas, a partir de los distintos “niveles” deseables que indica la literatura para las variables de estudio.
4. Establecer la estructura del modelo y evaluar su confiabilidad mediante la comparación con otros índices de calidad aceptados mundialmente y criterios técnicos.

RESUMEN DEL MARCO METODOLÓGICO

En esta sección se describen los criterios que se utilizaron en el proceso investigativo, sus características y fases principales para lograr los resultados esperados.

Características del estudio

El enfoque del estudio es cuantitativo debido a que las variables son cuantitativas y adquirieron sus valores de acuerdo con la información de los parámetros de monitoreo de la calidad del agua que realiza una institución privada de Guatemala en los ríos Xayá y Pixcayá.

El diseño adoptado es no experimental, puesto que los datos concernientes al monitoreo de la calidad del agua fueron analizados de forma directa, sin manipulación alguna. Además, es de tipo descriptivo ya que las pruebas estadísticas permitieron una descripción de la influencia de los parámetros de monitoreo en la calidad del agua y auxiliaron la generación de una propuesta de un modelo para estimar la calidad del agua.

El alcance del estudio es transversal ya que dicha base de datos corresponde a un período temporal específico (2007 – 2021).

Unidades de análisis

La población de estudio fue la base de datos de monitoreo de la calidad del agua de los ríos Xayá y Pixcayá elaborados por una institución privada de Guatemala durante los años 2007 a 2021.

Variables e indicadores

A continuación, se definen las variables de interés para el proceso investigativo.

Tabla 1.

Operativización de variables

Variable	Definición teórica	Definición operativa
x_i	Valor asociado al parámetro de la calidad del agua (demanda química de oxígeno, demanda bioquímica de oxígeno, sólidos totales disueltos, sólidos totales en suspensión, entre otros) que no ha sufrido ninguna transformación (anamorfismo), es decir, están en sus unidades habituales (mg/L, % m/v, entre otros). Estos se eligen a partir de la base de datos como los parámetros que se han medido en todos los años del período seleccionado.	Su valor es observacional, de acuerdo con los datos consignados por la institución privada. Las variables son de tipo cuantitativas continuas y, depende del caso, pueden o no ser de razón.
x_{if}	Valor asociado a una observación del parámetro de calidad del agua que ha sido transformado logarítmicamente.	Se obtiene de la transformación $x_{if} = \log_{10}(x_i + 1)$
z_{if}	Valor asociado a una observación del parámetro de calidad del agua que ha pasado por una transformación logarítmica y posteriormente por una estandarización robusta.	Se obtiene de la transformación siguiente: $z_{if} = \frac{x_{if} - m_f}{s_f}$ Donde m_f y s_f son la media y la desviación media absoluta de los

Continuación de la tabla 1.

Variable	Definición teórica	Definición operativa
		valores transformados de la variable x_f $(x_{1f}, x_{2f}, \dots, x_{nf})$. Se utiliza para trabajar el ACP y el AF con una escala unificada.
s_i	Valor máximo de tolerancia para la variable x_i.	Se utiliza como valor de referencia en la transformación de escala de unidades de variable x_i a unidades de subíndice q_i.
q_i	Subíndice de calidad del agua. Son aquellos parámetros de la calidad del agua que ya han pasado por alguna transformación (no hace referencia a las transformaciones x_{if} ni z_{if}). Son adimensionales.	La transformación se elige considerando los fenómenos físicos, químicos y microbiológicos, según aplique, y su influencia en la calidad del agua asociada al recurso hídrico y el uso provisto como agua de consumo humano. Su valor va de cero a uno, el cual resulta de una transformación matemática del valor del parámetro. Son variables cuantitativas continuas de razón.
w_i	Peso asignado al subíndice, q_i, que determina la importancia relativa del subíndice con respecto a los otros subíndices. Su valor se elige a través del análisis multivariado.	El peso se obtiene a través de los análisis de componentes principales y de rotación completa de la varianza. Son variables cuantitativas continuas de razón.

Nota. Se presenta la forma en que las variables serán operadas. Elaboración propia, realizado con Word.

Fases del Estudio

El desarrollo del ICA se realizó en cinco etapas generales, a saber:

- Fase 1. Revisión documental

La primera fase del proyecto se orienta a la recolección de información a través de fuentes bibliográficas que sean útiles para el estudio, ya sean fuentes primarias, secundarias o terciarias.

Las fuentes de interés fueron aquellas que ejemplificaron el proceso de creación de índices de calidad del agua a través de técnicas estadísticas estrictas o mezclas de técnicas estadísticas con técnicas iterativas o de lógica difusa.

También fueron de interés aquellas fuentes que permitieron asignar los niveles "deseables" de las variables, para los distintos usos potenciales que pueda adquirir el agua.

- Fase 2. Recolección de la Información

Para la segunda fase se solicitó el traslado de la base de datos a la institución privada, bajo los criterios de confidencialidad aplicables.

- Fase 3. Análisis de la Información

Esta fase del estudio se procesaron los datos obtenidos de la fase 2. Para ello se realizaron tres fases críticas:

Reducción de dimensionalidad, selección de parámetros influyentes y asignación de pesos: a través del análisis de componentes principales y análisis factorial.

Transformación de escala: a través de diversas técnicas, con lo que se incluyeron los modelos de regresión, según el caso, y se tomó como base el

conocimiento e interpretación de los fenómenos físicos, químicos o microbiológicos, según aplique, y su influencia en la calidad del agua asociada al recurso hídrico y el uso provisto como agua de consumo humano.

Generación del modelo de agregación: por medio de la interpretación de la información obtenida del análisis de componentes principales y la exploración de la estructura de los datos.

- Fase 4: Interpretación de la Información

En esta fase se comprobó la eficiencia del índice a través del contraste de resultados con la legislación, otros índices y criterio técnico. Con lo anterior se concluyó con respecto a los resultados obtenidos.

- Fase 5. Redacción del informe final

En esta fase se redactó el informe final, se incluyeron los resultados obtenidos, su discusión e interpretación, conclusiones y recomendaciones.

INTRODUCCIÓN

Este estudio es una sistematización que tiene como objetivo proveer una herramienta que permita estimar la calidad del agua de los ríos Xayá y Pixcayá mediante un número conocido como “índice”.

Los recursos hídricos del país se encuentran deteriorados y sufren de deterioro, esta problemática ha desencadenado ya una coyuntura social a nivel nacional y, además, demandas internacionales.

El desconocimiento de las implicaciones económicas, sociales y sanitarias que acarrea la contaminación de los cuerpos de agua, aunado a la falta de voluntad política, provoca que la contaminación en los cuerpos de agua de Guatemala aumente; inclusive los ríos Xayá y Pixcayá que abastecen entre el 30 y 40 % del agua potable de la región metropolitana de Guatemala.

Por otro lado, el lenguaje técnico y especializado con respecto a la calidad del agua dificulta la concienciación de las personas que no tienen una preparación académica en Biología, Microbiología y Química.

Guatemala no cuenta con herramienta de fácil interpretación que permita divulgar esta información a toda la población y la opción de “adoptar” un índice externo resulta compleja puesto que este responde a las condiciones del lugar para el cual fue creado.

Una opción viable para solventar esta situación es la construcción de un índice, el cual puede ser construido con el uso de técnicas estadísticas de análisis multivariado.

Este índice busca expresar la calidad del cuerpo de agua mediante un número que se asemeja a la calificación de una clase o de un examen (p. ej., 0 – 100). Debido a que incluso aquellas personas que han cursado el nivel primario están acostumbradas a esta expresión, el índice se perfila como una herramienta de divulgación científica comprensible por toda la población.

El acceso a la información científica es tan solo el beneficio más obvio de todos, son los impactos que se derivan en el tiempo los que cobran mayor relevancia. En orden de aparición se puede listar: comprensión, concienciación y sensibilización, reducción de contaminación, mayor valor inmediato del recurso hídrico, menor uso de recursos para saneamiento y menor riesgo asociado a la salud.

Se obtuvo un modelo para estimar la calidad del agua que se compone de siete variables seleccionadas y valoradas mediante análisis de componentes principales y análisis factorial. El modelo propuesto, denominado “ICAXP23”, fue comparado con los índices propuestos por Dinius en 1972-1987 y se concluyó que mostró un desempeño adecuado. Con ello se da por cumplido el objetivo general de este estudio y se genera una herramienta para la divulgación de información científica en una manera sencilla, al alcance de comprensión de todo tipo de público.

El enfoque adoptado en esta investigación fue cuantitativo, con diseño no experimental de tipo descriptivo y alcance transversal. Para ello, el estudio se dividió en cuatro etapas.

La primera consistió en la revisión exhaustiva de trabajos de investigación relacionados al desarrollo de índices de calidad del agua por técnicas estadísticas.

La segunda en la selección de variables, asignación de pesos y agregación. Para lo cual, se aplicaron las técnicas de análisis de componentes principales y factorial a una base de datos del monitoreo de la cuenca Xayá – Pixcayá, llevada a cabo durante los años 2007 a 2021 por una entidad privada que reclamó anonimato.

La tercera etapa consistió en la transformación de escala de las variables a una escala unitaria. Los datos utilizados fueron obtenidos de la revisión exhaustiva de reglamentación y criterios asociados a las variables influyentes.

La última etapa consistió en la revisión del desempeño del índice por comparación con otros índices, en la que se utilizó una muestra de las observaciones en la base de datos.

Debido a que la adecuación del muestreo para las técnicas de análisis de componentes principales y factorial no fue la esperada, la validez interna del estudio se vio comprometida. En cuanto a la validez externa, se sabe que un índice no puede ser aplicado a otros cuerpos de agua de manera indiscriminada, por lo que no se recomienda.

Este estudio se desarrolla en nueve capítulos, que se plantean a continuación:

El capítulo uno de este estudio presenta el marco referencial, en el cual se plasman los estudios previos que se consideraron para la elaboración de este trabajo de investigación.

En el capítulo dos se presenta el marco teórico, el cual recoge todos los fundamentos teóricos asociados a las técnicas estadísticas, los índices de calidad del agua y la base legal y científica concerniente a la calidad del agua.

Por su parte, el capítulo tres expone los resultados obtenidos de la aplicación de las técnicas de análisis de componentes principales, análisis factorial, transformaciones a escalas unitarias y evaluación del desempeño.

El capítulo cuatro presenta una discusión de los resultados obtenidos, expresada, tanto en términos de su validez interna, como externa.

1. MARCO REFERENCIAL

Un índice es un agregado de subíndices mediante una regla de agregación, el cual tiene como objeto resumir una gran cantidad de información asociada a un fenómeno, en este caso la calidad del agua, en un valor numérico único similar a una nota de examen.

Para construir un índice, habitualmente se contemplan cuatro etapas dispuestas en el siguiente orden: selección de variables, transformación de la escala asociada a las variables, asignación de pesos y agregación. Sin embargo, al utilizar las técnicas estadísticas multivariantes estas etapas no se delimitan de manera evidente ni necesariamente se realizan en el mismo orden.

La base de datos que se utilizó para este estudio fue suministrada por una entidad privada, por tanto, era imperativo realizar un análisis exploratorio, imputación de datos, observaciones, entre otros, para que el análisis multivariante entregara resultados confiables.

Yang, *et al*., (2020) descartaron las variables que se encontraban por debajo del límite de detección porque un valor tipificado como cero o “menor a” carece de sentido para el análisis de componentes principales. Este criterio fue utilizado y ampliado al límite superior para descartar las variables que no sirven para el ACP; específicamente cuando una gran cantidad de valores se reportan como “menor a” o “mayor a”.

Coletti, *et al*., (2010) iniciaron su estudio con un análisis exploratorio de los datos para evaluar el comportamiento de las variables y una matriz de

correlaciones de Spearman para evaluar relaciones subyacentes en el conjunto de variables. Estas técnicas fueron utilizadas para realizar inferencias iniciales de agrupamiento y relaciones subyacentes de las variables.

Debido a que múltiples autores hacen uso de las técnicas de ACP y AF para la selección de variables y la agregación de subíndices, este fue el método seleccionado. Cabe destacar que cada autor hace uso de la técnica con distintas transformaciones de datos, criterios de corte y técnicas de rotación, por ello, en los párrafos siguientes se describe más a detalle qué se recogió de cada estudio.

La adecuación del muestreo, multicolinealidad e interdependencia de los datos que se someten a ACP y AF deben ser verificadas. Así, la adecuación del muestreo se verificó mediante la prueba de Kaiser-Meyer-Olkin (KMO), utilizada tanto por Yang, como por Shirodkar. Mientras que, la interdependencia de las variables se verifica la prueba de Barlett, utilizada por Yang.

Otro factor de importancia para la aplicación de ACP es que las variables deben estar en la misma escala. Qian, *et al.*, (2007) abordan esta unificación de escala a través de una transformación logarítmica de los datos y su posterior estandarización por medio de una variante de la ecuación tradicional que implica utilizar la desviación media absoluta; lo que permite reducir el efecto de valores extremos. Debido a la presencia de valores extremos en la base de datos, se ha utilizado la transformación de escala descrita.

Los criterios de corte para definir el número de componentes principales a retener varían de autor en autor. Se comparó el criterio compuesto de Tripathi y Singal (2019) que considera retener los CCPP que cumplan con: valor propio mayor o igual a la unidad, varianza explicada por el componente mayor o igual a 0.10 (10 %) y varianza acumulada de los componentes mayor o igual a 0.60

(60 %) con otros criterios de corte y se demostró que un criterio de corte de $\lambda_0 = 0.7$ abordaba de mejor manera el problema.

Shirodkar, *et al*., (2010) hacen una contribución importante en su estudio, indicando que las cargas asociadas a cada variable en los componentes principales pueden categorizarse como fuertes (> 0.75), moderadas (0.5 – 0.75) y débiles (< 0.5). Su utilidad en la interpretación se utilizó tanto en los componentes principales, como en los componentes rotados.

Tripathi y Singal (2019) citan el método B4 propuesto por Jolliffe en 1972 para el descarte de variables y, consecuentemente, la retención de las variables influyentes. A raíz de dicha investigación se probó el método B4 de descarte de variables, pero se evidenció que se acomodaba mejor el método B2 para los fines de este estudio.

En cuanto al método de rotación para el AF, se considera que la selección de *Varimax* apunta a obtener una estructura más sencilla de interpretar. (Coletti, *et al.,* 2010; Flaieh, Mohammed-Ridha y Abdul-Ahad, 2014). Por su parte, Shirodkar, *et al*., (2010) indican que dicho método distribuye las cargas en los componentes principales de manera que su dispersión se maximiza al minimizar la cantidad de coeficientes grandes y pequeños y Tripathi y Singal (2019) enfatizan en que utilizarla tiene como consecuencia un menor número de coeficientes grandes en el mismo componente principal. Por lo anterior, se ha decidido utilizar el método de rotación *Varimax* para el AF.

Estos dos últimos autores hacen uso de los datos del AF para obtener los pesos asociados a cada parámetro, para ello elevaron los factores al cuadrado y los escalaron a la unidad. Este acercamiento se utilizó para asignar los pesos de las variables.

En este punto ya se ha realizado la selección de parámetros y la asignación de pesos. Además, se ha avanzado con la elucidación de la estructura de agregación. Continúa entonces el desarrollo del estudio con la transformación de escalas.

Flaieh, Mohammed-Ridha y Abdul-Ahad (2014) utilizaron un valor de tolerancia s_i para transformar la escala mediante una relación simple, c_i/s_i, en donde c_i es el valor asociado a una observación de la variable. La transformación utilizada en este estudio es una variante de la propuesta por los autores; permitiendo la interpretación de la escala en dirección menor a mayor calidad y con un rango acotado para múltiples fines provistos.

Se exploraron modelos aditivos-ponderados y multiplicativos-ponderados. Debido a ello se buscaron índices con estas características para poder comparar y que sus variables estuvieran contenidas en la base de datos del monitoreo, así los índices aditivo-ponderado y multiplicativo-ponderado propuestos por Dinius de 1972 a 1987 fueron utilizados para evaluar el desempeño del índice propuesto.

2. MARCO TEÓRICO

2.1. Fundamentos matemáticos y estadísticos

En esta sección se describirán los conceptos, métodos y técnicas matemáticas y estadísticas necesarias para abordar el problema de investigación, iniciará con los conceptos matemáticos de los que se auxilian las técnicas estadísticas por utilizar, continuará con los conceptos básicos estadísticos y finalizará con los conceptos y técnicas estadísticas avanzadas por utilizar.

2.1.1. Álgebra lineal

El álgebra lineal es una rama de la matemática que tiene como principal objetivo resolver ecuaciones lineales. Para ello se vale de vectores, matrices, espacios vectoriales y transformaciones lineales.

Sharipov (1996) señala que un enfoque para enunciar el álgebra lineal y la geometría multidimensional es el de coordenadas y matrices. Este enfoque, comienza con los sistemas de ecuaciones algebraicas lineales, luego se desarrolla la teoría de los determinantes, se considera el álgebra matricial y la geometría del espacio $\mathbb{R}^n$.

La estadística multivariada se vale del álgebra lineal para resolver problemas. El análisis multivariante es, de hecho, una aplicación del álgebra lineal en la estadística.

"La información estadística en AM es de carácter multidimensional, por lo tanto, la geometría, el cálculo matricial y las distribuciones multivariantes juegan un papel fundamental" (Cuadras, 2014, p. 13).

2.1.1.1. Vectores

Grossman y Flores (2012) identifican dos vectores: el vector renglón (ecuación 1) y el vector columna (ecuación 2). Tanto uno como el otro son conjuntos ordenados de n números ($x_1, x_2, \ldots$) y difieren únicamente en la forma en que se escriben sus componentes ($x_1, x_2, \ldots$).

$$\vec{r} = [x_1, x_2, \ldots, x_i] \quad (1)$$

$$\vec{k} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \end{bmatrix} \quad (2)$$

Un caso especial de vectores se da cuando todos sus componentes son cero, a este vector se le conoce como vector cero.

2.1.1.2. Matrices

Grossman y Flores (2012) definen a una matriz $m \times n$ como un arreglo rectangular de i, j números distribuidos en un orden de m renglones y n columnas y se simboliza con una letra mayúscula:

$$\mathrm{X} = [x_{ij}] = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1j} \\ x_{21} & x_{22} & \ldots & x_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \ldots & x_{ij} \end{bmatrix} \quad (3)$$

Así mismo identifican varios casos "especiales" de las matrices que, al igual que con los vectores, tienen especial importancia:

- Matriz cuadrada: es aquella matriz que tiene igual número de reglones que de columnas, es decir $m = n$.
- Matriz cero: es en la que todas sus entradas (o componentes) son cero.
- Matriz de un renglón ($1 \times n$): la cual genera un vector renglón.
- Matriz de una columna ($m \times 1$): la cual genera un vector columna.

Algunas matrices de especial importancia como las llamadas triangulares, escalar y la identidad se describirán en la próxima sección para comprensión de su utilidad.

2.1.1.3. Operaciones de vectores y matrices

Las operaciones matemáticas de los vectores y matrices difieren ligeramente del álgebra convencional.

A continuación, se hará una síntesis de las operaciones básicas según lo considerado por Grossman y Flores (2012):

La suma de matrices, $A + B$, sencillamente es la suma sus componentes. Así, sean $A = [a_{ij}]$ y $B = [b_{ij}]$, entonces $A + B$ se denota en la ecuación 4.

$$\mathrm{A} + \mathrm{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1j} + b_{1j} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2j} + b_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1} + b_{i1} & a_{i2} + b_{i2} & \dots & a_{ij} + b_{ij} \end{bmatrix} \quad (4)$$

De lo anterior se evidencia que, para que las matrices sean conformables para la suma, estas deben de ser de este orden.

La suma $A + B$ es conmutativa y asociativa. Además, su elemento neutro es la matriz cero $\mathbf{0}$, de este orden; la cual también recibe el nombre de matriz identidad para la suma.

El producto escalar se define como la multiplicación de un vector, $\vec{\mathrm{a}}$, o una matriz, $A = [a_{ij}]$, por un escalar, α, se define en las ecuaciones 5, 6 y 7.

$$\alpha\vec{a} = [\alpha a_1, \alpha a_2, \dots, \alpha a_i] \tag{5}$$

$$\alpha\vec{a} = \begin{bmatrix} \alpha a_1 \\ \alpha a_2 \\ \vdots \\ \alpha a_j \end{bmatrix} \tag{6}$$

$$\alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \dots & \alpha a_{1j} \\ \alpha a_{21} & \alpha a_{22} & \dots & \alpha a_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha a_{i1} & \alpha a_{i2} & \dots & \alpha a_{ij} \end{bmatrix} \tag{7}$$

La multiplicación descrita en las ecuaciones 5 a la 7 es distributiva y asociativa. Además, su elemento neutro es, como es de esperar, el escalar uno.

El producto interno (o producto punto) de dos vectores se simboliza con un punto, ·, y genera un escalar. El producto interno se define en la ecuación 8.

$$\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \tag{8}$$

Es importante notar que, en la ecuación 8, los vectores tienen la misma cantidad de componentes.

El producto vectorial (o producto cruz) de dos vectores, $\vec{a} = [\mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3]$ y $\vec{b} = [\mathrm{b}_1, \mathrm{b}_2, \mathrm{b}_3]$, se simboliza por una cruz, ×, y genera otro vector $\vec{c}$, como se establece en la ecuación 9.

$$\vec{a} \times \vec{b} = [a_2 b_3 - b_2 a_3, \quad a_1 b_3 - b_1 a_3, \quad a_1 b_2 - b_1 a_2] \tag{9}$$

Resulta importante notar que para la ecuación 9 solamente se define un vector en tres dimensiones. Aunque es posible generalizar el producto cruz a n dimensiones utilizando $n-1$ vectores, no se abordará en este documento.

Ahora bien, dadas dos matrices, $A = [a_{ij}]$ y $B = [b_{ij}]$, de dimensiones $m \times n$ y $n \times p$, su producto se obtiene a través del producto punto del renglón i y la columna j, genera una matriz de entradas c_{ij}, como se expresa en las ecuaciones 10 y 11.

$$c_{ij} = a_{i1}b_{ij} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} \tag{10}$$

$$\mathrm{A} \times B = \mathrm{C} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1j} \\ c_{21} & c_{22} & \dots & c_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ c_{i1} & c_{i2} & \dots & c_{ij} \end{bmatrix} \tag{11}$$

Nótese la importancia de la dimensión n, las matrices deben de coincidir en esta dimensión y en el orden de su operación para poder realizar el producto punto, es decir, para que el producto de dos matrices, $A \times B$, esté definido, la cantidad de columnas de A debe de coincidir con la cantidad de renglones de B.

Es importante resaltar que aún si $m = p$, el producto $A \times B$ no es necesariamente igual al producto $B \times A$ y por tanto la situación $A \times B = B \times A$ se considera como un caso especial.

Aunque la multiplicación de matrices no es necesariamente conmutativa, sí es asociativa, es decir que $A(B \times C) = (A \times B)C$ y también es distributiva con respecto a la adición, a saber $A(B + C) = A \times B + A \times C$.

Otro considerando en el producto de matrices es que $A \times B = 0$ puede obtenerse aun cuando $A \neq 0$ y $B \neq 0$. En este escenario se dice que A y B son

divisores de cero. Merece la pena aclarar que también existen casos en los que, para $A \times B = 0$, al menos una de las matrices es cero.

La multiplicación matricial también tiene un elemento neutro. El elemento neutro en este caso es una matriz escalar uno, con dimensiones que dependen del orden en que se plantee la multiplicación y de las dimensiones de la matriz A; como se evidencia en la ecuación 12.

$$I_m \times A = A \times I_n = A \tag{12}$$

Entonces, según la ecuación 13, al tener el elemento neutro se puede conseguir el inverso B de una matriz de una matriz cuadrada A, y si existe, es único.

$$A \times B = B \times A = I_n \tag{13}$$

Finalmente, en lo concerniente a la multiplicación de matrices, existen situaciones en las que $A \times B = A \times C$, aun cuando $B \neq C$. Por lo que no se debe concluir con ligereza que $B = C$, al efectuar una simple eliminación de la matriz A en ambos lados. Una última operación matricial que se debe de considerar es la transposición. Esta operación consiste en intercambiar los renglones y las columnas. Así, si una matriz A tiene orden $m \times n$, su transpuesta A^T tiene orden $n \times m$.

Algunas consideraciones importantes de la transposición de matrices se ejemplifican en las ecuaciones 14 y 15.

$$(A + B + \cdots + N)^T = A^T + B^T + \cdots + N^T \tag{14}$$

$$(A \times B)^T = B^T \times A^T \tag{15}$$

2.1.2. Conceptos estadísticos básicos

En esta sección se describirán los conceptos estadísticos básicos para la aplicación posterior de las técnicas estadísticas avanzadas, a saber: el análisis multivariante (AEM) y la regresión lineal.

2.1.2.1. Medidas de localización

Las medidas de localización, también conocidas como medidas de tendencia central (MTC), están diseñadas según Walpole, *et al*., (2012) para brindar a la persona que analiza los datos, valores cuantitativos de la ubicación central de las observaciones asociadas a una muestra. Por su parte Triola (2018) define las MTC como "un valor medio o en el centro de un conjunto de datos" (p. 82). Las medidas de localización o posición que se desarrollarán en este apartado son la media ($\bar{x}$), la mediana ($\tilde{x}$) y la moda, como medidas básicas, y las medias geométrica y armónica, como medidas complementarias.

Según estos mismos autores, la media es un promedio numérico de un conjunto de datos, así, si se tienen las observaciones muestrales $x_1, x_2, \dots, x_n$, entonces la media muestral se simboliza como $\bar{x}$ y se define matemáticamente como sigue:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{16}$$

"La mediana tiene el propósito de reflejar la tendencia central de la muestra sin que sea afectada por valores extremos" (Walpole, *et al*., 2012, p. 11). Para obtener la mediana, los valores de las observaciones deben de ordenarse en forma creciente. Así, la mediana se denota como $\tilde{x}$, "M" o "Med" y define según la ecuación 17 si es impar y según la ecuación 18 si es par.

$$\tilde{x} = x_{(n+1)/2} \tag{17}$$

$$\tilde{x} = \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) \quad (18)$$

Triola (2018) define:

> La moda como el valor que ocurre con mayor frecuencia y resalta que, dado conjunto de datos, pueden existir varios valores que empaten en frecuencia o incluso que ningún valor se repita por lo que, un grupo de datos puede ser unimodal, bimodal, multimodal o puede no tener moda (p. 85).

Además, señala que la moda no suele utilizarse en datos cuantitativos sino en cualitativos.

La media armónica, simbolizada como H, "se utiliza a menudo como una medida de tendencia central para conjuntos de datos que consisten en tasas de cambio, como velocidades" (Triola, 2018, p. 96). Spiegel y Stephens (2009) definen la media armónica como el recíproco de la media aritmética de los recíprocos de los números (ver ecuación 19).

$$\mathrm{H} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \quad (19)$$

La media geométrica, que se simboliza como G, "se utiliza a menudo en los negocios y la economía para encontrar tasas de cambio promedio, tasas de crecimiento promedio o razones medias" (Triola, 2018, p. 96). Spiegel y Stephens (2009) por su parte definen la media geométrica como la raíz enésima del producto de los números (ver ecuación 20).

$$G = (x_1 x_2 \ldots x_n)^{1/n} \quad (20)$$

2.1.2.2. Medidas de variabilidad

Spiegel y Stephens (2009) indican que “el grado de dispersión de los datos numéricos respecto a un valor promedio se llama dispersión o variación de los datos” (p. 95).

Además, Triola (2018) señala que:

> La variación es el tema más importante de la estadística. En esta sección se definirán las medidas más representativas de dispersión. Para empezar, se resalta la importancia del redondeo en las medidas de dispersión. Indica que, al redondear el valor de una medida de variación, conserve un decimal más que los presentes en el conjunto original de datos. (p. 97)

El rango es una medida asociada a la dispersión de los datos que Triola (2018) define como “la diferencia entre el valor máximo de datos y el valor mínimo de datos” (p. 98). Al ser sensible a valores extremos, no es una media robusta de variabilidad.

Walpole, *et al.,* (2012) señalan que “la medida muestral de dispersión que se utiliza más a menudo es la desviación estándar de la muestra” (p. 15). No obstante, resulta importante definir antes a la varianza para entender por qué dicha medida es un reflejo de la variabilidad en un grupo de datos. La varianza muestral se define en la ecuación 21.

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad (21)$$

De la ecuación 21 se puede notar que, a medida que los valores se alejan de la media muestral, la varianza aumenta. Entonces, se puede conceptualizar la desviación estándar muestral como la raíz cuadrada de la varianza muestral (ecuación 22); de allí que esta sea una medida de variabilidad.

$$s = \sqrt{s^2} \tag{22}$$

Entonces, según Triola (2018), la desviación estándar muestral "es una medida de cuánto se desvían los valores de datos de la media" (p. 98).

Otras medidas de variabilidad que son de utilidad se pueden estimar a por medio de los cuantiles. En este aspecto, son de especial importancia los percentiles 10 (decil 1), 25 (cuartil 1), 75 (cuartil 3) y 90 (decil 9). Las medidas de dispersión asociadas son el rango intercuartílico, el rango semi intercuartílico y el rango percentil.

Según Spiegel y Stephens (2009), los cuartiles dividen al conjunto de datos en cuatro partes iguales y se simbolizan como Q_1, Q_2 y Q_3. Los deciles por su parte dividen al conjunto de datos en diez partes iguales y se simbolizan como $D_1, D_2, \dots, D_9$. Finalmente, los percentiles dividen el conjunto de datos en cien partes iguales y se simbolizan como $P_1, P_2, \dots, P_{99}$.

Dicho lo anterior, "el rango intercuartílico es la diferencia entre Q_3 y Q_1, el rango semi intercuartílico es la mitad del rango intercuartílico y el rango percentil es la diferencia entre P_{90} y P_{10}. Otros rangos pueden utilizarse, pero no es común" (Spiegel y Stephens, 2009, p. 95-96). Los valores que se obtienen de estos rangos dan una idea de cómo se dispersan los datos en los distintos intervalos.

2.1.2.3. Distribuciones continuas de probabilidad

En esta sección se describirán algunas de las distribuciones continuas de probabilidad que presentan utilidad o pueden ser útiles, para resolver el problema de investigación. Entre ellas: la distribución normal (o *Gaussiana*) y las distribuciones muestrales asociadas a la media y a la varianza.

Las distribuciones continuas de probabilidad indican la probabilidad con la que una variable aleatoria continua puede asumir ciertos valores.

> Además, Triola (2018) señala que: la gráfica de cualquier distribución de probabilidad continua se denomina curva de densidad y cualquier curva de densidad debe satisfacer el requisito de que el área total bajo la curva es exactamente 1. Por ello, concluye que existe una correspondencia entre el área bajo la curva y la probabilidad. (p. 229)

La primera, y la más importante, es la distribución normal, la cual, según Walpole, *et al*., (2012), "describe de manera aproximada muchos fenómenos que ocurren en la naturaleza, la industria y la investigación. Además, los errores en las mediciones científicas se aproximan muy bien mediante una distribución normal" (p. 172).

"Una variable aleatoria continua que sigue una distribución normal, descrita por la ecuación 23, se denomina variable aleatoria normal" (Walpole, *et al*., 2012, p. 173).

$$n(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty \qquad (23)$$

De la ecuación 23 se puede observar que para definir una curva normal se deben de conocer la media (μ) y la desviación estándar (σ). Por tanto, “hay una cantidad infinita de distribuciones normales diferentes, depende de los valores utilizados para la media y la desviación estándar” (Triola, 2008, p. 228).

Con el fin de facilitar el trabajo con cualquier distribución normal no estándar, primero se “estandariza” la distribución con auxilio de la ecuación 24.

$$z = \frac{x-\mu}{\sigma} \qquad (24)$$

En la ecuación 24, x y z son ambas variables aleatorias normales, la distinción radica en que para x su media y su desviación estándar pueden variar, pero para z, son necesariamente cero y uno. Por ello, a z se le conoce como variable aleatoria normal estándar.

Según Walpole, *et al*., (2012), dada la variable x, su función de densidad se define según la ecuación 23. Debido a que esta se puede transformar a z utilizando la ecuación 24, uno se puede valer del cálculo diferencial e integral para obtener el área bajo la curva entre un par de puntos x_1, x_2. En consecuencia, se puede establecer la probabilidad de hallar a la variable entre dichos puntos. Los cálculos asociados no se detallarán en este documento.

Hasta ahora se ha abordado la distribución normal a partir de que se conocen la media poblacional y la desviación estándar poblacional, pero usualmente no es así. Las distribuciones muestrales son importantes en la estadística ya que sirven para hacer inferencias con respecto a μ y σ.

Para iniciar a comprender las distribuciones muestrales se debe de conocer el teorema del límite central, el cual permite hacer inferencias sobre la media.

Walpole, *et al.* (2012), lo definen así:

Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n, tomada de una población con media μ y varianza finita σ^2, entonces la forma límite de la distribución de

$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \quad (25)$$

a medida que $n \to \infty$, es la distribución normal estándar $n(z; 0,1)$.

La aproximación normal para $\bar{X}$ por lo general será buena si $n \geq 30$, siempre y cuando la distribución de la población no sea muy asimétrica. Si $n < 30$, la aproximación será buena sólo si la población no es muy diferente de una distribución normal y, como antes se estableció, si se sabe que la población es normal, la distribución muestral de $\bar{X}$ seguirá siendo una distribución normal exacta, sin importar qué tan pequeño sea el tamaño de las muestras. (p. 234)

La segunda distribución muestral en este apartado es la de la varianza. Walpole, *et al*., (2012) indican que, si de una población normal se extrae una muestra aleatoria y cuantifica su varianza se obtiene el estadístico s^2.

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{(n-1)s^2}{\sigma^2} + \frac{(\bar{x}-\mu)^2}{\sigma^2/n} \quad (26)$$

La variable aleatoria $\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$ tiene una distribución ji-cuadrada y n grados de libertad. La deducción de la ecuación siguiente no se abordará, pero se concluye que, si s^2 es la varianza asociada a una muestra aleatoria que se sustrajo de una población normal con varianza σ^2, entonces:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} \quad (27)$$

Tiene una distribución ji-cuadrada con $n-1$ grados de libertad. Nótese que un reacomodo rápido de los términos genera la ecuación 27.

Finalmente, dada una población, si se desconoce su varianza no es posible realizar inferencias inmediatas utilizando el TLC. La distribución T (o t de Student) toma en consideración el desconocimiento de la varianza poblacional para realizar inferencias concernientes a la media. De manera que:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad (28)$$

Nótese que se reemplazó σ por s de la ecuación 25 a la ecuación 28. La función de densidad para dicha distribución está dada por:

$$h(t) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi\nu}} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} \quad (29)$$

En donde $-\infty < t < \infty +$, con $n-1$ grados de libertad.

2.1.3. Regresión lineal

Al intentar describir la relación existente entre varias variables existen dos escenarios puntuales, el primero es que la relación entre la variable sea exacta y por tanto determinística (p. ej., el volumen de un cubo está determinado por el largo de sus caras mediante la ecuación $V = l^3$), el segundo surge de cuando la relación no es exacta y contiene una componente probabilística. (Walpole, *et al*., 2012, 389).

De igual manera señalan que el análisis de regresión sirve como herramienta para encontrar la mejor relación entre las variables.

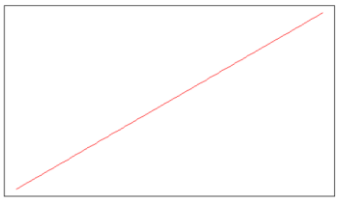
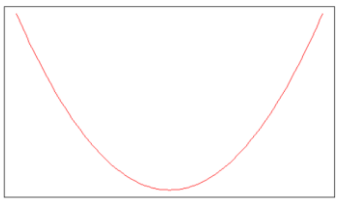
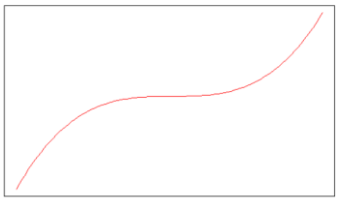
2.1.3.1. Curvas de aproximación

A diferencia de muchos autores, Spiegel y Stephens (2009) abordan el proceso inicial de una manera más intuitiva que la de iniciar con diagramas de dispersión y continuar directamente con estadística. En este aspecto, los autores introducen el término “curva de aproximación” para incluir el proceso intuitivo antes de la estadística.

> Para identificar la curva de aproximación, proponen utilizar funciones matemáticas con estructuras bien conocidas y luego identificar el patrón en el gráfico de dispersión. En la tabla 1 se muestran algunas curvas de aproximación polinómicas. Otras funciones matemáticas como la hipérbola, curva exponencial, curva geométrica, curva de Gompertz y curva logística pueden ser utilizadas como curvas de aproximación. (Spiegel y Stephens, 2009, p. 318)

Tabla 2.

Funciones matemáticas para curvas de aproximación

Nombre de la función	Función	Representación gráfica
Polinomio de primer grado	$Y = \alpha_0 + \alpha_1 x$	
Polinomio de segundo grado	$Y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$	
Polinomio de tercer grado	$Y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$	

Nota. En la tabla se presentan las funciones matemáticas utilizadas. Elaboración propia, realizado con R Studio.

2.1.3.2. Modelo de regresión lineal

"La estructura de las funciones matemáticas en la tabla 2 es determinista. Cuando la correspondencia entre las variables no es determinista se debe de introducir un componente aleatorio" (Walpole, *et al*., 2012, p. 390). A partir de

este momento se utilizará la ecuación de la recta (ecuación 30) y se cambiará la notación utilizada para introducir el componente aleatorio (ecuación 31).

$$Y = \alpha_0 + \alpha_1 x \tag{30}$$

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{x} + \epsilon \tag{31}$$

Merece la pena ahondar en la nomenclatura de la ecuación 31, ya que β_0 y β_1 corresponden a la intersección y la pendiente, respectivamente, pero, además, son también los parámetros "reales". Por otro lado, ϵ es una variable aleatoria, denominada "error aleatorio", que cumple con $E(\epsilon) = 0$ y $V(\epsilon) = \sigma^2$ (Walpole, *et al*., 2012, p. 391).

El error aleatorio, ϵ, está dado por:

$$y_i - E(Y_i) = y_i - (\beta_0 + \beta_1 x_i) = \epsilon_i \tag{32}$$

La ecuación 32 resulta aleccionadora puesto que, como indican Walpole, *et al*., (2012), es de esperar que los pares observados (x_i, y_i) se encuentren dispersos alrededor de la recta definida por $E(Y) = \beta_0 + \beta_1 x$ y que el error individual sea dado por $\epsilon_i = y_i - E(Y_i)$, con una varianza homogénea para cumplir con los supuestos del error aleatorio.

En la práctica, los coeficientes de la ecuación 32 se estiman a partir de datos muestrales por lo que conviene reescribir dicha ecuación como sigue:

$$y = b_0 + b_1 x + e \tag{33}$$

La naturaleza muestral de la ecuación 33 y la necesidad de estimar β_0, β_1 en la ecuación 31 con un nivel de confianza dado, hace evidente la posibilidad de generar inferencias derivadas del análisis de regresión lineal.

Finalmente, es conveniente recordar que, aunque este apartado y los subsiguientes se basan, por facilidad, en la ecuación de la recta para exponer los conceptos; los modelos lineales no son los únicos y por eso el término “curva de ajuste resulta útil para fines generales mientras que el término recta de ajuste hace alusión únicamente al modelo de la recta” (Spiegel y Stephens, 2009, 319).

2.1.3.3. Método de mínimos cuadrados

El método de mínimos cuadrados sirve de herramienta para evitar el empleo del criterio personal y se emplea para estimar los coeficientes de una curva de ajuste (Spiegel y Stephens, 2009, 319).

Además, Walpole, *et al.*, (2012) señalan que:

> Es frecuente que se utilice el modelo de regresión lineal simple aun cuando se sepa que el modelo no es lineal o que se desconozca la estructura verdadera. Este método suele ser acertado, en particular cuando el rango de las x es estrecho. (p. 414)

Para entender la técnica de mínimos cuadrados se debe de definir qué es un residual.

En este aspecto, se define como un error ajuste del modelo $\widehat{y_i} = b_0 + b_1 x$, según la ecuación 34.

$$e_i = y_i - \widehat{y_i} = y_i - (b_0 + b_1 x_i) \qquad (34)$$

La idea detrás del modelo es calcular b_0 y b_1 “de manera tal que la suma de los cuadrados de los residuales sea mínima” (Walpole, *et al*., 2012, p. 395).

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \quad (35)$$

Al resultado de la ecuación 35 se le conoce como suma de cuadrados del error (SCE), con respecto a la recta de ajuste.

“Las ecuaciones para calcular b_0 y b_1 se obtienen por diferenciación de b_0 y b_1 en la ecuación 35” (Walpole, *et al.,* 2012, p. 396). El proceso de diferenciación no se abordará, pero las ecuaciones resultantes son

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad (36)$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad (37)$$

Con las ecuaciones 36 y 37 ya se pueden calcular b_0 y b_1 para generar la recta de ajuste de la ecuación 33. Sin embargo, estos autores indican que, para poder hacer inferencias sobre los parámetros β_0 y β_1 de la ecuación 31, se debe de conocer la varianza relacionada a los estimadores. Para ello, se introduce la siguiente ecuación.

$$s^2 = \frac{SCE}{n-2} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2} \quad (38)$$

2.1.3.4. Inferencias sobre los coeficientes

Las primeras dos inferencias de importancia que se pueden hacer con respecto a β_0 y β_1 son sus intervalos de confianza (IC). Según lo indican Walpole, *et al*., (2012), los IC para β_0 y β_1 se calculan mediante las ecuaciones 39 y 40.

$$b_1 - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} \quad (39)$$

$$b_0 - t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2} < \beta_0 < b_0 + t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2} \quad (40)$$

Para el caso de la pendiente, es importante probar que difiere de cero. Esto porque si $\beta = 0$ entonces no hay relación lineal entre ambas variables.

Generalizan la prueba con la siguiente hipótesis:

$$\begin{cases} H_0: & \beta_1 = \beta_{10} \\ H_1: & \beta_1 \neq \beta_{10} \end{cases}$$

En donde β_{10} es un valor para poner a prueba la hipótesis mediante el estadístico:

$$t = \frac{b_1 - \beta_{10}}{s/\sqrt{S_{xx}}} \quad (41)$$

2.1.3.5. Coeficientes de correlación y determinación

Según Triola (2018):

El coeficiente de correlación lineal r mide la fuerza de la correlación lineal entre los valores cuantitativos pareados x y y en una muestra. El coeficiente de correlación lineal (conocido también como coeficiente de correlación del momento del producto de Pearson) asume valores de -1 a +1, incluidos ambos extremos (p. 472).

Dicho coeficiente se calcula según la ecuación 42.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \quad (42)$$

Triola (2018) indica que “el diseño de r permite estimar la fuerza de la relación lineal, lo cual limita su aplicación a dicho modelo” (p. 473).

Además, resalta que la correlación no es un indicador de causalidad y que los datos basados en promedios aumentan el valor del coeficiente porque suprimen la variación individual.

“Otro coeficiente de utilidad se conoce como coeficiente de determinación, R^2. A diferencia del coeficiente de correlación, el coeficiente de determinación es un indicador de la proporción de la variabilidad explicada por el modelo de ajuste; en forma de proporción” (Walpole, *et al*., 2012, p. 407). Dicho coeficiente se calcula como sigue

$$r^2 = 1 - \frac{SCE}{STCC} \quad (43)$$

Triola (2018) resalta que el referido coeficiente se puede calcular por medio de elevar el coeficiente de correlación al cuadrado y sugiere utilizar dicha técnica.

Tanto la utilización del coeficiente de correlación como del coeficiente de determinación como único criterio para poner modelos en competencia y posteriormente tomar una decisión de selección no es recomendable.

“El coeficiente de determinación presenta tendencia alcista cuando se agregan variables regresoras, es decir, por el de sobre ajustar el modelo” (Walpole, *et al*., 2012, 408). Esto adquiere especial importancia cuando se realiza regresión lineal múltiple.

"Como respuesta al comportamiento alcista de r^2 cuando se agregan variables al modelo, se utiliza una versión conocida como coeficiente de determinación ajustado, r_{aj}^2, el cual toma en cuenta tanto la cantidad de variables regresoras, k, como el tamaño de la muestra" (Triola, 2018, p. 513).

$$r_{aj}^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1 - r^2) \quad (44)$$

2.1.4. Análisis estadístico multivariado

El análisis estadístico multivariado, al cual se hará referencia de aquí en adelante como "AEM", es la colección de técnicas estadísticas enfocadas a estudiar, analizar e interpretar los datos que se obtenidos de observar p variables en n individuos.

Cuadras (2014) señala que en AEM la información es multidimensional y que en consecuencia tanto la geometría, como el álgebra lineal y las distribuciones multivariantes son fundamentales.

En esta sección se abordarán los fundamentos estadísticos del AEM y las técnicas de análisis de componentes principales (ACP), análisis de factores (AF) y clasificación jerárquica o análisis de *clusters*.

2.1.4.1. Matrices de datos

Para iniciar, se debe de tener claro que la información en AEM se presenta en matrices de datos, como las que se definieron en la sección 2.1.1.2 mediante la ecuación 3. En este aspecto, la información en la matriz X, de dimensiones $n \times p$ se dispone de manera que los n individuos se encuentren en las filas y las p variables en las columnas.

Cuadras (2014) propone la ecuación 45 como representación matricial de los datos:

$$\mathrm{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \tag{45}$$

En adelante se debe de tener claro que x_i es la i-ésima fila de X y que x_j es la j-ésima columna de X.

2.1.4.2. Análisis de componentes principales

Se iniciará este apartado indicando el objetivo del ACP que, según Härdle y Simar (2015) indican, es reducir la dimensionalidad.

Lo anterior se entiende si se contempla a cada una de las p variables consideradas para ACP como una dimensión y que, la variable compuesta (o latente) que surge de la combinación lineal de estas es una dimensión subyacente que explica una proporción de la variabilidad total; poniendo un límite mínimo de explicación de la variabilidad se puede seleccionar un grupo de las variables compuestas.

Las variables compuestas se conocen como componentes principales (CCPP), la definición que Cuadras (2014) propone para los componentes principales es: “Las componentes principales son unas variables compuestas incorrelacionadas tales que unas pocas explican la mayor parte de variabilidad de X” (p. 77).

El proceso para llevar a cabo un ACP se resume en trasladar los datos a una matriz como en la ecuación 45, centrar los datos, generar una matriz de covarianzas (S), generar una matriz de correlaciones (R), obtener los eigenvalores (λ_i), obtener los eigenvectores (γ_i), calcular los CCPP, seleccionar los CCPP a retener e interpretar los CCPP.

Para la mayoría de estas etapas Cuadras (2014) indica que existen expresiones matriciales sencillas, las cuales se presentan a continuación:

Vector de medias ($\bar{x}$):

$$\bar{x}' = \frac{1}{n} 1'X \tag{46}$$

Matriz de centrado (H):

$$\mathrm{H} = \mathrm{I} - \frac{I}{n} 11' \tag{47}$$

Donde 1 es el vector de unos de dimensiones $n \times 1$.

Matriz de datos centrados ($\bar{X}$):

$$\bar{X} = X - 1\bar{x}' = HX \tag{48}$$

Matriz de covarianzas (S):

$$S = \frac{1}{n} \bar{X}'\bar{X} = \frac{1}{n} X'HX \tag{49}$$

Matriz de correlaciones (R):

$$R = D^{-1/2} S D^{-1/2} \tag{50}$$

Donde $D^{-1/2}$ es la matriz diagonal con los inversos de las desviaciones de las variables.

No obstante, no pasa lo mismo con los valores y vectores propios (λ_i, γ_i), en estos los cálculos se vuelven más complejos a medida que incrementa el número de variables; requiriendo de *softwares* especializados para poder llevar a cabo estos cálculos.

Dicho lo anterior, no se ahondará más en este apartado en los conceptos matemáticos en los que se basa la técnica de ACP, sino en sus consideraciones e interpretaciones prácticas.

En materia de los supuestos que se deben cumplir para llevar a cabo el ACP, no figuran los supuestos de normalidad como uno tendería a pensar, en cambio, la escala de las variables juega un papel fundamental.

Según Cuadras (2014), el ACP y AF son técnicas con un fundamento geométrico mediante el cual buscan similitudes o disimilitudes a través de conceptos de distancias. Además, Härdle y Simar (2003) indican que el ACP y el AF no son invariantes a la escala.

Por lo tanto, se requiere que las variables se encuentren en una escala común. La transformación de escala depende de cada caso, pero algunas transformaciones de escala comunes son la logarítmica y el valor z (*z-score*).

Seguido de esto, se debe considerar que las variables no deben presentar multicolinealidad, pero deben presentar interdependencia. Las pruebas de Haitovsky y de Bartlett permiten dar respuesta a si se cumplen estos supuestos o no.

La prueba de Haitovsky tiene por hipótesis nula que la determinante de la matriz de correlaciones es cero y es descrita por Rockwell (1975) en la ecuación (51). Tiene una distribución ji-cuadrado con $\nu = p(p-1)$ grados de libertad.

$$\chi_h^2(v) = \left(1 + \frac{2p+5}{6} - n\right) ln(1 - \det((R)) \quad (51)$$

Donde:

p: número de variables

n: número de observaciones

R: matriz de correlaciones

"La prueba de Bartlett, también llamada de esfericidad, independencia completa o de interdependencia, es una prueba de utilidad en ACP, mediante la cual se plantea la hipótesis nula de que las p variables son estocásticamente independientes" (Cuadras, 2014, p. 85).

$$H_0: \Sigma = \Sigma_d = \text{diag}(\sigma_{11}, \dots, \sigma_{pp}) \quad (\mu \text{ desconocida}) \quad (52)$$

El estadístico de prueba se da en la ecuación 53:

$$-2log\lambda_r = -nlog|R| \sim \chi_q^2 \quad (53)$$

Un último supuesto debe tomarse en consideración y es el valor de la prueba propuesta por Kaiser (1970), conocida a menudo como prueba de Kaiser-Meyer-Olkin (KMO). El valor constituye una medida de la adecuación del muestreo y según el autor parece responder a cuatro efectos: el número de variables, de observaciones, de factores latentes y a la correlación general $\bar{r}$.

Aunque es difícil señalar alguno en específico, Kaiser y Rice (1974) proponen una escala en la que un valor de prueba por debajo de 0.50 es inaceptable (aunque muchos autores toman un valor de 0.60).

Los CCPP pueden estimarse, tanto de la matriz de correlaciones, R, como de la matriz de varianzas-covarianzas (S). La elección depende de cómo se

encuentren las variables, si se encuentran en la misma escala se utiliza la matriz S, por el contrario, si están en distintas escalas se utiliza la matriz R.

Los CCPP guardan ciertas propiedades: ortogonalidad, varianza máxima y son variables compuestas (combinaciones lineales) de las variables observables.

Una variable compuesta, Y, es una combinación lineal de las variables observables con coeficientes $a = (a_1, \dots, a_p)'$ (Cuadras, 2014).

$$Y = a_1X_1 + \cdots + a_pX_p = Xa \tag{54}$$

En ACP, la transformación de p variables observables genera p variables compuestas (o CCPP). Así, la ecuación 54 se puede transformar a la expresión generalizada:

$$Y_i = \sum_{j=1}^{p} a_{ij}X_j \tag{55}$$

Donde los coeficientes a_{ij} son el peso de la variable j, en el CCPP i.

La construcción de los CCPP se realiza de tal manera que Y_1 retiene la máxima varianza del conjunto de datos, Y_2 retiene la máxima varianza no retenida por Y_1, y así consecutivamente; permitiendo su orden y priorización.

Además, su construcción debe asegurar la ortogonalidad entre los distintos pares de Y_i, por lo que los coeficientes cumplen con:

$$a_{i1}a_{j1} + a_{i2}a_{j2} + \cdots + a_{ip}a_{jp} \qquad \forall\, i \neq j \tag{56}$$

Finalmente, se limita la escala mediante la condición:

$$a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1 \qquad i = 1,2, \dots, p \tag{57}$$

Lo que sigue es encontrar los valores propios (*eigenvalores*) y vectores propios (*eigenvectores*) mediante paquetes estadísticos. Con estos, se pueden obtener los CCPP. A los coeficientes a_{ij} de los CCPP también se les llama carga factorial o puntaje y permiten hacer inferencias considerando la importancia de la variable en el CCPP.

Según el autor anteriormente mencionado, la varianza de Y_i está dada por $var(Y_i) = \lambda_i$ y la variación total es $tr(S) = \sum_{i=1}^{p} \lambda_i$. Por lo que, el porcentaje de variabilidad explicada por las m primeras CCPP está dado por:

$$P_m = 100 \frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p} \tag{58}$$

Existen diversos métodos para escoger la cantidad adecuada de componentes principales. A continuación, se exponen dos de ellos:

El primero consiste en establecer un porcentaje mínimo P de manera arbitraria, así si se establece que $P = 80$ y $P_m > 80$, se ha cumplido con el criterio del porcentaje arbitrario con m componentes principales.

El segundo criterio se conoce como “criterio de Kaiser” y consiste en retener las m primeras componentes tales que $\lambda_m \geq 1$. Aunque estudios Montecarlo indican que el punto de corte (λ_0) es realmente $\lambda^* \geq 0.7$ (Cuadras, 2014, p. 86-88)

2.1.4.3. Análisis factorial

La segunda técnica de AEM que se abordará es el análisis de factores (AF). Cuadras (2014) indica que: “El análisis factorial es un método multivariante que pretende expresar p variables observables como una combinación lineal de m variables hipotéticas o latentes, denominadas factores” (p. 97).

> Aunque la formulación de la técnica de AF se parece a la de ACP, su modelo es distinto. Inicia con el planteamiento más simple de un factor común, se puede plantear que dadas $X_1, \dots, X_p$ variables observables, la covariabilidad de estas se recoge en un factor común F y se asignan p factores únicos $U_1, \dots, U_p$, uno para cada variable, los cuales aportan la variabilidad no explicada por el factor común. (Cuadras, 2014, p. 78)

Así, el modelo factorial se describe a continuación:

$$X_i = a_i F + d_i U_{i,} \quad i = 1, \dots, p \tag{59}$$

Este modelo supone que las variables y los factores están estandarizados y que, además, los $p + 1$ factores no se encuentran correlacionados.

> Especial importancia adquiere el coeficiente a_i, conocido como coeficiente de saturación de la variable X_i en el factor F; el cual denota la correlación entre X_i y el factor común. Además, al cuadrado del coeficiente de saturación, a_i^2, se le llama comunalidad y expresa la variabilidad explicada por el factor F. (Cuadras, 2014, p. 98)

Cuadras (2014) señala que para que el modelo de un factor sea válido se debe cumplir con las relaciones tetrádicas, cumpliendo con que las relaciones entre los elementos de la misma columna no diagonal de dos filas de la matriz de correlaciones sean constantes.

Sin embargo, en la práctica usualmente se requiere de un modelo con múltiples factores y por tanto el modelo planteado en la ecuación 59 se escala a:

$$X_1 = a_{11}F_1 + \cdots + a_{1m}F_m + d_1U_1 \tag{60}$$

$$X_2 = a_{21}F_1 + \cdots + a_{2m}F_m + d_2U_2$$

$$X_p = a_{p1}F_1 + \cdots + a_{pm}F_m + d_pU_p$$

El modelo tiene tres supuestos, el primero es que los factores comunes y los factores únicos no están correlacionados dos a dos.

$$\mathrm{cor}\left(F_i, F_j\right) = 0, \quad i \neq j = 1, \dots, m \tag{61}$$

$$\mathrm{cor}\left(U_i, U_j\right) = 0, \quad i \neq j = 1, \dots, p \tag{62}$$

Además, supone que los factores comunes no se relacionan con los factores únicos.

$$\mathrm{cor}\left(F_i, U_j\right) = 0, \quad \mathrm{i} = 1, \dots, \mathrm{m}, \quad \mathrm{j} = 1, \dots, \mathrm{p} \tag{63}$$

El tercero es que todos los factores son variables reducidas de media cero y varianza uno.

Este mismo autor señala que el AF tiene como objetivo encontrar e interpretar la matriz factorial. Por tanto, para aplicar la técnica se debe de definir la matriz factorial, A, que contiene los coeficientes de saturación, a_{ij}.

$$\mathrm{A} = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ a_{21} & \dots & a_{2m} \\ \vdots & \vdots & \vdots \\ a_{p1} & \dots & a_{pm} \end{bmatrix} \tag{64}$$

También indica que, si se considera el vector de variables como $X = (X_1, \dots, X_p)'$, el vector de los factores comunes $F = (F_1, \dots, F_p)'$, y al vector de los factores individuales $U = (U_1, \dots, U_p)'$, entonces la expresión matricial del modelo factorial está dada por:

$$X = AF + DU \tag{65}$$

El número máximo de factores comunes, m, está limitado por la expresión siguiente:

$$\mathrm{m} \leq \mathrm{m_a} = 1/2(2\mathrm{p} + 1 - \sqrt{8\mathrm{p} + 1}\,) \tag{66}$$

2.2. Índices de calidad o contaminación del agua

En esta sección se abordarán las generalidades de los índices necesarias para su construcción. No se ahondará en ejemplos puntuales de creación de subíndices más allá de lo estrictamente necesario. Usualmente se designan las siglas ICA para referirse a un índice de calidad del agua y las siglas ICO para un índice de contaminación del agua.

2.2.1. Definición

Abbasi y Abbasi (2012) definen los índices como sigue:

> Se puede decir que los índices (el singular de los índices es índice) son representaciones compuestas de una condición o situación derivada de una combinación, hecha de cierta manera, de varios hechos/mediciones observadas relevantes, pero no proporcionales. La combinación conduce a un único número ordinal que facilita la comprensión e interpretación de la importancia general de los hechos que han contribuido a ese número. (pág. 4)

Calvo-Brenes (2018) enfatiza que, para que un índice sea práctico, debe de reducir una cantidad grande de indicadores a una forma más simple. Y León (1992) indica que un buen diseño del índice hace que el valor único que este genera sea representativo de la calidad del agua.

Un índice (I) es un agregado de subíndices (s_i) mediante una regla de agregación (Agg), a saber:

$$I = Agg(s_1, s_2, \dots, s_N) \tag{67}$$

Por lo tanto, un índice es una integración de los valores asociados a los parámetros más influyentes de forma que se genere un valor único, adimensional, en una escala comprensible. Además, este valor único tiene utilidad en tanto que el índice sea diseñado de manera adecuada.

2.2.2. Características

Según Abbasi y Abbasi (2012), los índices tienen cuatro características inherentes:

- Ambigüedad: se da cuando el índice excede el valor crítico sin que ninguno de los subíndices supera su valor crítico asociado.
- Eclipsado: se da si el índice no excede el valor crítico aun cuando uno o más de los subíndices supera su valor crítico.
- Compensación: se da cuando el índice no está desviado hacia los extremos de ambigüedad o eclipsado, atributo que se da cuando se crean índices libres de alguno de los extremos. Así, se dice que un índice compensado se da cuando:
- $min_{i=1}^{N}(s_i) \leq Agg(s_1, s_2, \dots, s_N) = I \leq max_{i=1}^{N}(s_i)$ (68)
- Rigidez: se presenta cuando se requiere introducir subíndices adicionales al índice o cuando se utiliza el índice en un lugar distinto al que fue creado pero sus características ambientales difieren de manera significativa y el modelo de agregación no lo permite. Esto se da principalmente en modelos de tipo multiplicativo.

2.2.3. Tipos de índices

Los índices se pueden categorizar de varias maneras y es importante conocer su clasificación debido a que de esto dependen sus propiedades, alcance e interpretación.

2.2.3.1. Por el significado de la escala

Calvo-Brenes (2018) indica que los índices pueden elaborarse a partir de dos enfoques diferentes:

- Índices de contaminación (ICO): aquellos en los cuales el número del índice se incrementa con los niveles crecientes de contaminación.

- Índices de calidad (ICA): aquellos en los cuales el número del índice decrece con los niveles crecientes de contaminación.

No obstante, algunos índices no encajan dentro de esta definición. En el siguiente enunciado se propone un enfoque más generalizado con el afán de incluir las posibles transformaciones matemáticas: un índice de contaminación es aquel que expresa el grado de contaminación y por tanto su valor es proporcional a la contaminación y un índice de calidad es aquel que expresa el grado de calidad y por tanto su valor es inversamente proporcional a la contaminación (proporcional a la calidad).

2.2.3.2. Por la metodología utilizada para su creación

Según la metodología utilizada para su construcción Abbasi y Abbasi (2012) clasifican los índices como convencionales, de lógica difusa, estadísticos o estocásticos. En las líneas siguientes se describen los tres grupos:

- Índices creados con técnicas convencionales: son aquellos que se crean por medio de métodos ajenos a la estadística avanzada, inteligencia artificial o teoría probabilística. Por ejemplo, métodos iterativos como el método de Delphi.
- Índices creados con lógica difusa: son aquellos que hacen uso del conocimiento vago del fenómeno y el uso de la inteligencia artificial. Así, una computadora puede programarse con el conocimiento del fenómeno y producir, mediante algoritmos de distintos grados de complejidad, inferencias con respecto al fenómeno. Por ejemplo, si x y y entonces z.

- Índices creados con estadística avanzada: como su nombre lo indica son aquellos que se obtienen a través del uso de técnicas estadísticas avanzadas (análisis multivariante, análisis de series de tiempo, otras).

2.2.3.3. Por sus componentes

Un índice también puede clasificarse de acuerdo con sus constituyentes en cuatro grupos:

- Biológicos o bióticos: son aquellos que toman únicamente factores bióticos pluricelulares como indicadores (p. ej., a través del estudio de comunidades de peces, plantas acuáticas, macroinvertebrados bénticos, plancton y otros).
- Microbiológicos: son aquellos que toman únicamente factores bióticos unicelulares (p. ej., enterobacterias, coliformes totales, fecales, *E. coli*, entre otros).
- Fisicoquímicos: según Bhargava (1983) son en los que se consideran únicamente factores físicos (p. ej., temperatura, conductividad, turbiedad, entre otros) o químicos (fósforo, nitrógeno, metales pesados, entre otros).
- Mixtos: aquellos que incluyen dos o más de las categorías anteriores.

2.2.3.4. Por su regla de agregación

Abbasi y Abbasi (2012) señalan que muchos métodos de agregación de subíndices son posibles y que los tres más básicos son:

- Aditivo: cuando la operación básica para la agregación es la suma (ecuación 69)

$$I = \sum_{i=1}^{n} s_i w_i \quad (69)$$

- Multiplicativo: cuando la operación básica para la agregación es la multiplicación (ecuación 70)

$$I = \prod_{i=1}^{n} s_i^{w_i} \quad (70)$$

- Lógico: cuando el cálculo se opera a través de un proceso lógico del tipo "si x y y entonces z", como en el caso de índices de mínimos y máximos (ecuaciones 71 y 72)

$$I = min(s_i) \quad (71)$$

$$I = max(s_i) \quad (72)$$

2.2.4. Creación

Diversos acercamientos se han realizado para la creación de los ICA, desde técnicas básicas iterativas hasta la utilización de técnicas estadísticas avanzadas.

Sea cual sea la metodología que se utilice para su creación Abbasi y Abbasi (2012) divide el proceso en cuatro etapas generales, selección de parámetros, transformación de parámetros, asignación de pesos y agregación de subíndices.

2.2.4.1. Selección de parámetros

No existe una regla para seleccionar los parámetros que constituirán los subíndices. Su selección ha dependido de la metodología utilizada para la creación del índice y de los participantes de dicho proceso. Sin embargo, Barker y Kramer utilizaron seis criterios para seleccionar los parámetros:

- Que sean amplia y regularmente medidos
- Que tengan efectos la vida acuática, el uso recreacional del agua o en ambos.
- Que provengan de fuentes antropogénicas
- Que sean controlables mediante programas de reducción de contaminantes.
- Que tengan rangos realistas (desde no contaminación hasta contaminación grave).
- Que sean sensibles a pequeños cambios en la calidad del agua

De los criterios anteriores, únicamente el segundo se compone de enfoques puntuales por lo que, si se generalizara su redacción a que tengan efectos claves en las actividades o usos de interés; se podrían adoptar estos para la selección de los parámetros relevantes en las actividades o usos definidos.

Calvo-Brenes (2018) agrega que se deben de evitar los indicadores que son redundantes entre sí. Condición que puede ser evaluada por análisis estadísticos de correlación.

Finalmente, Abbasi y Abbasi (2012) señalan que, “la selección de los parámetros es crucial para la utilidad del índice, pero usualmente está llena de subjetividad e incertidumbre” (p.11). Por ello, cada vez es más común su construcción con acercamientos estadísticos que intentan reducir la incertidumbre y subjetividad del proceso.

2.2.4.2. Transformación de parámetros

Abbasi y Abbasi (2012) indican que los parámetros elegidos deben de transformarse a una escala en común ya que: usualmente se expresan en distintas unidades, los rangos en los cuales se presentan varían ampliamente o tienen distintos comportamientos en términos de concentración – impacto.

Generalizan la construcción de los subíndices (s_i) considerando que la transformación de un parámetro (x_i) depende de una función (f_i), lo cual queda representado en la ecuación 73:

$$s_i = f_i(x_i) \tag{73}$$

Además, también distinguen cuatro tipos de transformaciones, la primera, la transformación lineal se representa en la ecuación 74:

$$s_i = \beta_0 + \beta_1 x_i \tag{74}$$

La segunda se deriva de la primera y se da cuando la transformación se compone de dos o más líneas rectas, tiene especial importancia la función escalonada con puntos de quiebre $(a_1, b_1), (a_2, b_2), \dots, (a_j, b_j)$ (ecuación 75)

$$s_i = \frac{b_{j+1} - b_j}{a_{j+1} - a_j}(x_i - a_j) + \mathrm{b_j}, \qquad \mathrm{a_j} \leq x_i \leq \mathrm{a_{j+1}} \tag{75}$$

La tercera se da cuando la transformación no es lineal, la cual puede tener asociada una función implícita o una función explícita (esta última se ejemplifica con la ecuación 76).

$$s_i = \beta_0 + x_i^{\beta_1} \tag{76}$$

La cuarta se deriva de la tercera y se da cuando la transformación, además de no ser lineal, está segmentada en varias funciones en las que, al menos una, no es lineal. Sin embargo, hasta aquí únicamente se han revisado las transformaciones propuestas y no cómo se llega a estas. Al igual que en todas las etapas de la creación de índices, diversos acercamientos se han llevado a cabo. Calvo-Brenes (2018) ejemplifica varios de estos acercamientos:

- A través de tablas: en el índice de Horton se realizaron tablas en las que se asignó un valor único al subíndice para un rango dado del parámetro.
- A través de gráficos: Brown *et al*., (1970, 1973) y Landwehr (1974), propusieron transformaciones a través de gráficas de calidad contra concentración y una función implícita asociada.
- A través de funciones: Walski y Parker (1974) presentaron transformaciones a través de funciones explícitas.

Resulta evidente que las transformaciones que dependen de funciones matemáticas explícitas tienen ventajas en materia de automatización por ordenadores; tanto para el cálculo del subíndice como para estimaciones de incertidumbre. Las transformaciones de este tipo, sin embargo, pueden ser más engorrosas de crear que las demás y pueden generar errores de estimación en puntos específicos si no se hacen de manera adecuada.

2.2.4.3. Asignación de pesos

La asignación de pesos es un proceso fuertemente criticado que no se observa en la totalidad de los índices puesto que algunos de ellos asumen igual importancia de cada uno de los subíndices.

En este aspecto Abbasi y Abbasi (2012) señalan:

> Aún después de haber reducido la lista a 10 – 20 parámetros, se requiere asignar pesos a cada parámetro. Porque incluso cuando todos los parámetros seleccionados son considerados como indicadores importantes de la calidad del agua, aún no son igualmente importantes. Entre los parámetros seleccionados, algunos serán de mayor importancia que algunos otros. (p. 15)

Por su parte, Calvo-Brenes (2018) indica:

> “En el caso de los índices elaborados para un uso específico del agua, la ponderación de los indicadores puede resultar conveniente. Sin embargo, cuando los índices evalúan la calidad en forma general, esta metodología puede resultar contraproducente.” (p. 271 – 272)

La subjetividad de este proceso es un tema de crítica frecuente, en el que se contemplan dos métodos utilizados con rutina para este proceso: las encuestas y el método de Delphi. El segundo surge como alternativa del primero en un intento de reducir la subjetividad (Abbasi y Abbasi, 2012; Calvo-Brenes, 2018).

No obstante, Calvo-Brenes (2018) también menciona la técnica de análisis de correlación bivariada de Pearson como una alternativa que no introduce un panel de expertos en el proceso y que reduce el tiempo necesario para la obtención de los pesos.

2.2.4.4. Agregación de subíndices

En la sección 2.2.3.4 se da una breve introducción a los tipos de agregación de los subíndices (aditivo, multiplicativo, lógico). Sin embargo, como Abbasi y Abbasi (2012) señalan que varios métodos de agregación son posibles.

A continuación, se describirán los métodos de agregación según lo que indican Abbasi y Abbasi (2012):

El primer método es la suma lineal, esta se calcula mediante la adición de los subíndices no ponderados como se indica en la ecuación 77.

$$I = \sum_{i=1}^{n} s_i \tag{77}$$

Un índice de suma lineal posee la desventaja de que presenta ambigüedad, puesto que, si hipotéticamente se genera el índice I, compuesto únicamente de los subíndices s_1 y s_2, en donde $s_i = 0$ es la ausencia del parámetro y $s_i = 100$ su valor de fallo; el índice podría reflejar el valor de fallo $I = 100$ aun cuando ninguno de los dos subíndices falle.

El segundo método es la suma ponderada, esta se calcula de manera similar al método de suma lineal, pero se introduce un parámetro adicional conocido como peso (w_i), que cumple con $\sum w_i = 1$, como se observa en la ecuación 78.

$$\sum_{i=1}^{n} s_i w_i \tag{78}$$

Un caso especial se da cuando $w_1 = w_2 = \cdots = w_n$, lo cual es una media aritmética.

Los índices de suma ponderada reducen el problema de ambigüedad, pero incurren en un problema de eclipsado dado que, al volver al caso hipotético anterior; el índice podría no reflejar el valor de fallo $I = 100$ aun cuando uno de los dos subíndices falle.

El tercer método es el de raíz – suma – potencia (*root sum power*, en inglés), el calcula mediante una conjunción de las tres operaciones y la incorporación del valor p, un número real positivo mayor que uno, según la ecuación 79.

$$I = [\sum_{i=1}^{n} {s_i}^p]^{1/p} \tag{79}$$

La técnica tiene la ventaja de no presentar ni ambigüedad ni eclipsado pero la condición limitante de la función la hace compleja de manipular.

El cuarto método es el multiplicativo, el cual se basa en la operación de multiplicación. Su forma más común en los ICA es el producto ponderado, según la ecuación 80.

$$I = \prod_{i=1}^{n} {s_i}^{w_i} \tag{80}$$

Al igual que se describió para el método de la suma ponderada, se debe de cumplir con $\sum w_i = 1$. En este caso, el índice será cero cuando al menos un subíndice sea cero. Además, cuando al menos un subíndice llegue al valor de fallo, el índice será sensible al valor de fallo. Por ello, este método reduce los problemas de ambigüedad y eclipsado.

Un caso especial de este método se da cuando $\mathrm{w} = w_1 = w_2 = \cdots = w_n$, en cuyo caso se obtiene una media geométrica según la ecuación 81.

$$I = [\prod_{i=1}^{n} s_i]^w = [\prod_{i=1}^{n} s_i]^{1/n} \quad (81)$$

El último método es la agregación lógica, en la cual se utilizan los operadores mínimo y máximo. Aunque estos operadores tienen la ventaja de que no existe la posibilidad de ambigüedad ni de eclipsado, su uso en índices de calidad ha sido limitado.

El operador mínimo se utiliza para subíndices con escala decreciente mientras que el operador máximo se utiliza con subíndices con escala creciente. La forma general que estos operadores adquieren se muestra en las ecuaciones 82 y 83.

$$I = min(s_1, s_2, \dots, s_n) \quad (82)$$

$$I = max(s_1, s_2, \dots, s_n) \quad (83)$$

2.3. Marco Legal

Son muy pocos los reglamentos, normativos y decretos que constituyen la base legal asociada al recurso hídrico en Guatemala. Eso desvela la subvaloración que se le da a dicho recurso vital y que se extiende hacia los recursos naturales que en su conjunto constituyen la base para un desarrollo sostenible de la sociedad.

Cabe resaltar que la Comisión Guatemalteca de Normas (COGUANOR) genera normas técnicas elegibles y no obligatorias, por lo que, para que estas normas tengan carácter obligatorio requieren que un Acuerdo Gubernativo, Acuerdo Ministerial, Decreto del Congreso o Decreto Municipal así lo indique.

Los demás compromisos ambientales se derivan de compromisos internacionales, principalmente aquellos que asume la iniciativa privada, y que derivan en sistemas de gestión, producción o monitoreo.

A continuación, se presentan los principales documentos que constituyen el marco legal asociado al recurso hídrico:

2.3.1. Constitución Política de la República de Guatemala

En 1985, por medio de la Asamblea Nacional Constituyente, se emitió la Constitución Política de la República de Guatemala que se encuentra vigente. Esta, en su Artículo 97, concerniente a medio ambiente y equilibrio ecológico, plantea que el Estado, las municipalidades y los habitantes del territorio nacional están obligados a propiciar el desarrollo social, económico y tecnológico que prevenga la contaminación del ambiente y mantenga el equilibrio ecológico. Se dictarán todas las normas necesarias para garantizar que la utilización y el aprovechamiento de la fauna, de la flora, de la tierra y del agua, se realicen racionalmente, evitando su depredación.

2.3.2. Ley de protección y mejoramiento del medio ambiente

En 1986 se creó la Ley de protección y mejoramiento del medio ambiente a través del Decreto del Congreso Número 68-86. Sin embargo, no fue sino hasta el año 2000 que se creó el Ministerio de Ambiente y Recursos Naturales (MARN) según el Artículo 5 del Decreto del Congreso Número 90-2000, el cual modificó el Artículo 2 del Decreto 68-86 expone que la aplicación de esta ley y de sus reglamentos compete al Organismo Ejecutivo, a través del Ministerio de Ambiente y Recursos Naturales, cuyas funciones establece la Ley del Organismo Ejecutivo.

2.3.3. Código de Salud

El código de salud, Decreto del Congreso Número 90-97, en su Artículo 38 indica que, en relación con el ambiente, las acciones de promoción y prevención buscarán el acceso de la población con énfasis en la de mayor postergación, a servicios de agua potable, adecuada eliminación y disposición de excretas, adecuada disposición de desechos sólidos, higiene de alimentos, disminución de la contaminación ambiental.

Es responsabilidad de las municipalidades la distribución del servicio del agua potable y de su saneamiento, según lo establece la Constitución Política de la República y el Código de Salud.

2.3.4. Agua para consumo humano

En el año 1999 se publicó la primera versión de la norma COGUANOR NTG 29001, la norma establece los valores críticos de las características que debe tener el agua para considerarse como apta para el consumo humano. Se divide en cuatro tipos de parámetros: características físicas y organolépticas, químicas, bacteriológicas y radiológicas. Para todos los analitos se definen límites máximos permisibles (arriba de los cuales existe daño a la salud del consumidor) y para algunos se definen también límites máximos admisibles (arriba de los cuales existe percepción sensorial pero no daño a la salud del consumidor).

2.3.5. Agua envasada para consumo humano

En el mismo año que se publicó la norma COGUANOR NGO 29001, se publicó la norma COGUANOR NGO 29005:99. De igual manera que la norma de agua potable, esta norma cuenta con las secciones de características físicas y organolépticas, químicas, bacteriológicas y radiológicas, pero le agrega dos numerales destinados a cloración y ozonificación. En esta norma también se definen límites máximos permisibles y para algunos se definen límites máximos admisibles.

2.3.6. Acuerdo Gubernativo 137-2006

En el 2006 se emite el reglamento de evaluación, control y seguimiento ambiental (RECSA), el objeto del reglamento se indica en el Artículo 1 que el presente Reglamento contiene los lineamientos, estructura y procedimientos necesarios para apoyar el desarrollo sostenible del país en el tema ambiental, estableciendo reglas para el uso de instrumentos y guías que faciliten la evaluación, control y seguimiento ambiental de los proyectos, obras, industrias o actividades, que se desarrollan y los que se pretenden desarrollar en el país.

En este se definen los mecanismos para la evaluación, control y seguimiento ambiental mediante estudios de impacto ambiental (EIA), otorgamiento de licencias ambientales y la definición de sanciones y multas.

Cabe destacar que el RECSA aplica a todo proyecto, obra, industria o actividad que, por sus características, pueden producir deterioro a los recursos naturales, renovables o no, al ambiente, o introducir modificaciones nocivas o notorias al paisaje y a los recursos culturales del patrimonio nacional; según lo indica en el Artículo 5.

2.3.7. Acuerdo Gubernativo 236-2006

En el 2006 se emite el reglamento de las descargas y reúso de aguas residuales y de la disposición de lodos y de la disposición de lodos, a través del Acuerdo Gubernativo 236-2006.

El Acuerdo Gubernativo establece un modelo de reducción progresiva de las cargas de demanda bioquímica de oxígeno en los efluentes y una serie de valores máximos para los otros dieciséis parámetros que componen este. En todos los casos los plazos establecidos para entes generadores ya existentes se previeron para finalizarse en el año 2024, con metas previstas cada cuatro años y con la primera meta prevista para el 2015.

2.3.8. Otra reglamentación

Muchos de los requerimientos ambientales, incluso el recurso hídrico, que se realizan en el país proviene de contratos internacionales de la industria privada. Por ejemplo, la Ley de Modernización de la Inocuidad Alimentaria de Estados Unidos (FSMA), disposiciones asociadas a residuos de plaguicidas y algunas otras como los convenios con bancos internacionales.

3. RESULTADOS

A continuación, se muestran los resultados obtenidos en el estudio. Tales resultados se agrupan en los distintos objetivos de este y se subdividen en las distintas etapas del análisis de los datos.

3.1. Objetivo 1. Seleccionar las variables influyentes mediante la reducción de la dimensionalidad, con las técnicas de análisis de componentes principales y análisis factorial

Para tal efecto se inicia con la exploración de la base de datos, la cual se presenta a continuación.

3.1.1. Exploración de la base de datos

La base de datos fue suministrada por una entidad privada que reclamó mantener su identidad en anonimato. Se recibió una base de datos precariamente elaborada, con múltiples falencias por lo que, previo a llevar a cabo el ACP se corrigieron estas.

3.1.1.1. Base de datos sin tratamiento

La exploración inicial de la base de datos reveló lo siguiente:

- Compuesta por 45 variables distribuidas en cinco puntos de muestreo y 17 fechas.

- Los límites inferiores y superiores no fueron definidos por la entidad o se careció de homogeneización de criterios en la elaboración de la base de datos.
- Las cifras significativas no fueron definidas por la entidad o se careció de criterios en la elaboración de la base de datos.
- Las fechas de toma de muestra no fueron las mismas entre los distintos puntos de muestreo.
- Las variables medidas variaron tanto en fecha como en lugar.

3.1.1.2. Base de datos estandarizada

Debido a las falencias encontradas, la estructura de la base de datos suministrada se estandarizó y se obtuvo el resumen que se muestra en la tabla 2.

Con base en dicho resumen, las variables con menor número de valores faltantes (n=18), con una escala cuantitativa y límites definidos fueron retenidas. Así, se retuvieron 14 variables, lo que representa únicamente un 31 % de las variables que la base de datos cruda contenía.

Las variables retenidas fueron: conductividad eléctrica (CE), calcio (Ca), cloruro (Cl), color (COL), dureza (DUR), hierro (Fe), magnesio (Mg), nitrito (NO2), nitrato (NO3), sólidos disueltos totales (SDT), sulfato (SO4), temperatura (TEM), turbidez (TUR) y potencial de hidronio (pH).

Los resultados que de este punto en adelante se muestran se basan únicamente en el conjunto de variables retenidas. Las fechas constituyeron las observaciones.

Tabla 3.

Resumen de variables y valores faltantes

Analito	N	n	p	Analito	N	n	p
B	85	84	0.9882	Alc. total	85	42	0.4941
PO4	85	83	0.9765	NH3	85	42	0.4941
Hg	85	82	0.9647	F	85	31	0.3647
Cu	85	81	0.9529	Mn	85	28	0.3294
Cd	85	79	0.9294	Aspecto	85	18	0.2118
SiO2	85	79	0.9294	C. fecales	85	18	0.2118
N	85	77	0.9059	C. totales	85	18	0.2118
P	85	77	0.9059	CE	85	18	0.2118
Pb	85	76	0.8941	Ca	85	18	0.2118
CN	85	74	0.8706	Cl	85	18	0.2118
Cr (VI)	85	71	0.8353	COL	85	18	0.2118
Fenoles	85	71	0.8353	DUR	85	18	0.2118
Na	85	71	0.8353	Fe	85	18	0.2118
As	85	70	0.8235	Mg	85	18	0.2118
K	85	69	0.8118	NO2	85	18	0.2118
OD	85	65	0.7647	NO3	85	18	0.2118
Detergentes	85	58	0.6824	Olor	85	18	0.2118
SF	85	55	0.6471	SDT	85	18	0.2118
ST	85	55	0.6471	SO4	85	18	0.2118
SV	85	55	0.6471	TEM	85	18	0.2118
SES	85	54	0.6353	TUR	85	18	0.2118
DBO	85	52	0.6118	pH	85	18	0.2118
DQO	85	46	0.5412				

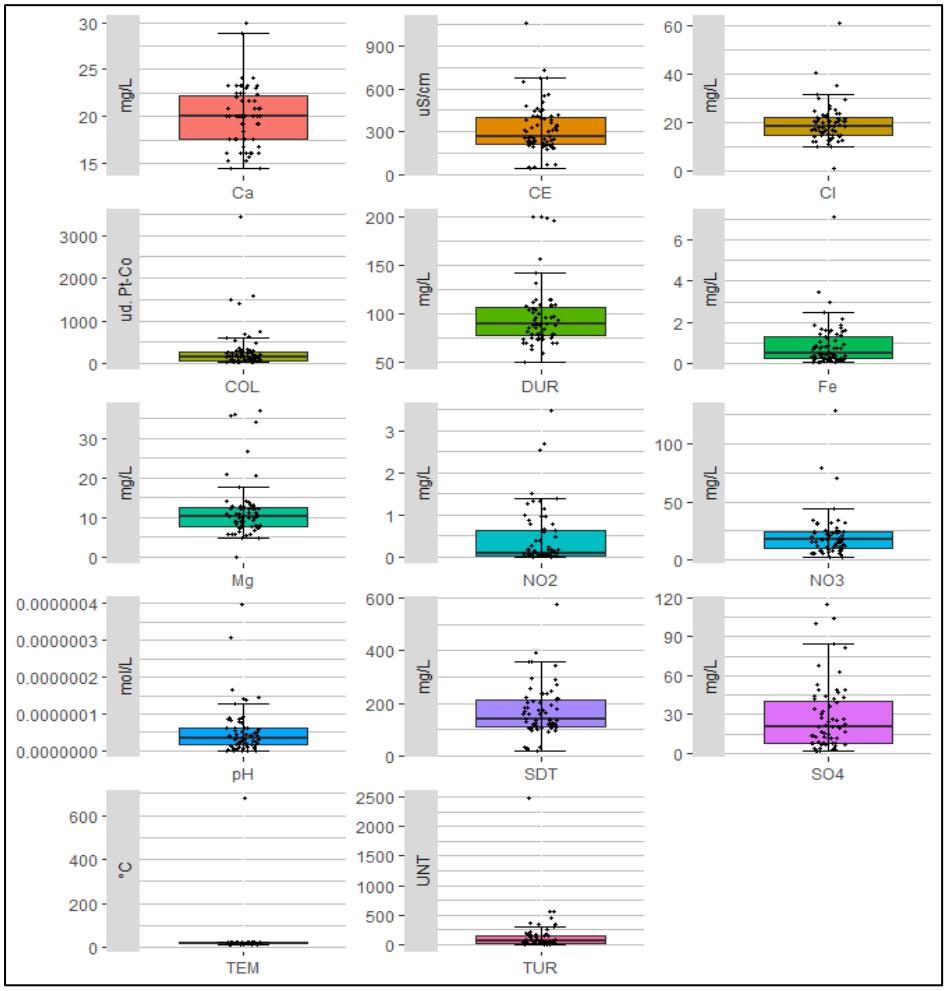
Nota. N: número total de observaciones, n: número de valores faltantes, p: proporción de valores faltantes. Elaboración propia, realizado con R Studio.

3.1.1.3. Estadística descriptiva

Se llevó a cabo una exploración visual de las variables (ver figura 1), encontrando valores atípicos en dureza (1.4 mg/L), hierro (104 mg/L), temperatura (678 °C), sulfato (374 mg/L) y turbidez (2 475 UNT). Tras las consideraciones técnicas, únicamente se imputaron los primeros tres.

Figura 1.

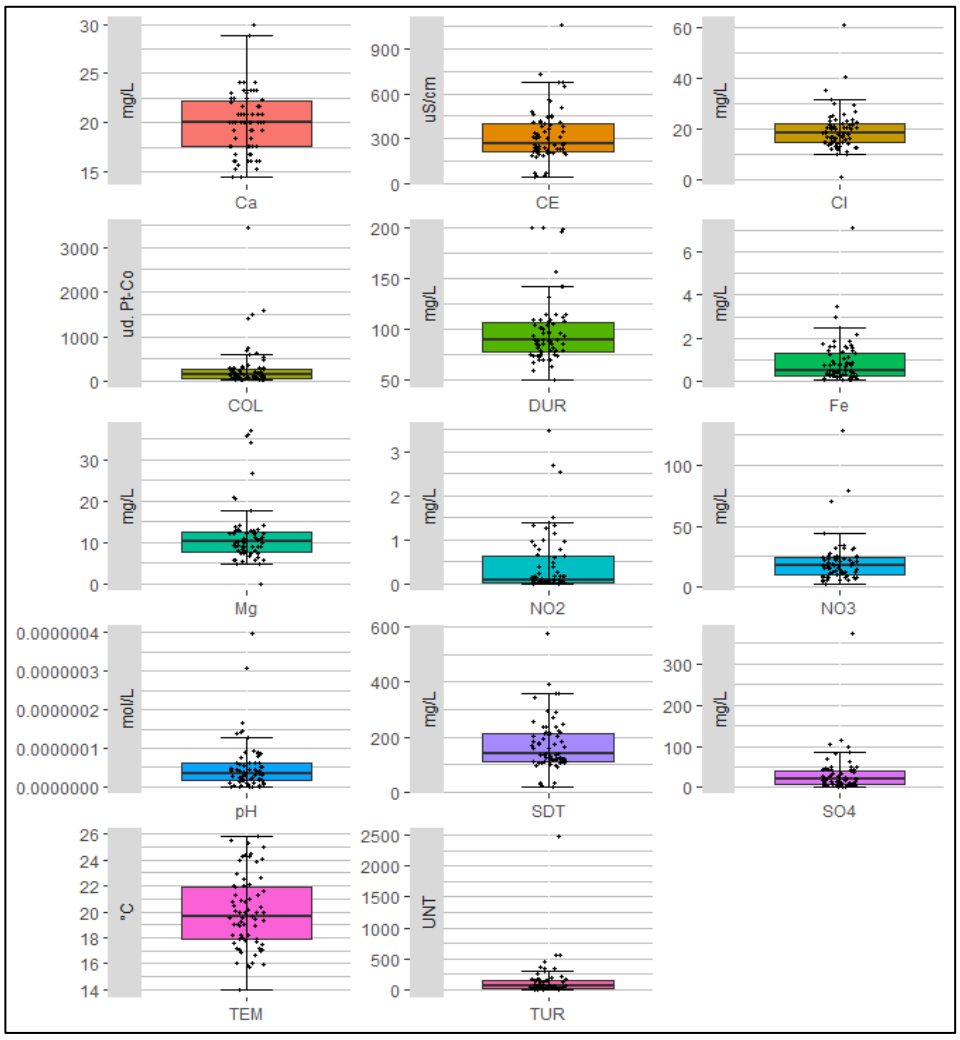
Gráficos de caja de las variables con valores atípicos



Nota. Exploración visual de las variables. Elaboración propia, realizado con R Studio.

Figura 2.

Gráficos de caja de las variables sin valores atípicos



Nota. Tras la imputación de valores atípicos, se obtuvieron los gráficos de caja y bigotes Elaboración propia, realizado con R Studio.

Tabla 4.

Estadísticos descriptivos

	n	$\bar{x}$	s	Mín.	Máx.	Rango	Asimetría	Curtosis	Unidades
Mg	67	11.73	7.276	0.00	37.05	37.05	2.2	4.7	mg/L
CE	67	320.58	175.800	40.00	1066.00	1026.00	1.5	3.5	uS/cm
NO2	67	0.42	0.689	0.00	3.48	3.47	2.4	6.2	mg N/L
Ca	67	20.01	3.119	14.42	30.00	15.58	0.5	0.7	ud. Pt-Co
Cl	67	19.69	8.123	1.10	61.50	60.40	2.2	9.1	mg/L
Color	67	297.99	503.237	18.00	3450.00	3432.00	4.3	21.5	mg/L
Dureza	67	97.60	32.146	50.00	200.00	150.00	1.8	3.3	mg/L
Fe	67	0.91	1.064	0.03	7.15	7.12	3.3	15.5	mg/L
NO3	67	20.48	18.869	3.00	128.60	125.60	3.5	15.4	mg N/L
SO4	67	31.79	49.628	2.00	374.00	372.00	5.0	31.1	ud. pH
Temperatura	67	19.98	2.817	14.00	25.80	11.80	0.3	-0.8	mg/L
Turbidez	67	143.65	315.735	3.57	2475.00	2471.43	6.1	41.5	mg/L
pH	67	7.58	0.611	6.40	9.32	2.92	0.9	0.7	°C
SDT	67	170.16	94.191	21.00	576.00	555.00	1.5	3.8	UNT

Nota*. El resumen de estadísticos descriptivos. Elaboración propia, realizado con R Studio.*

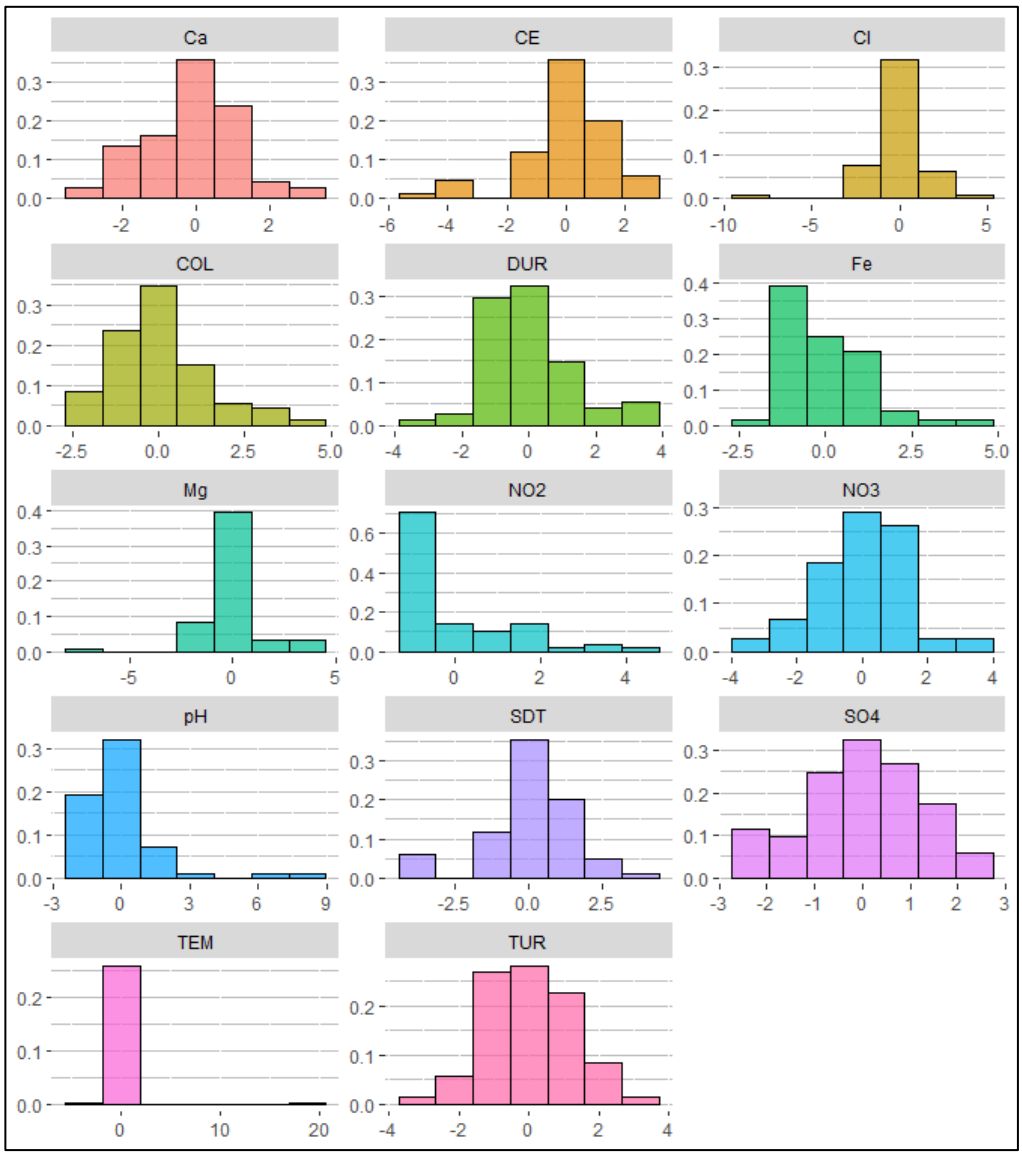
3.1.2. Estandarización de escala

Debido a que el ACP y el AF no son invariantes a la escala, se debe guardar especial cuidado en que todas las variables se encuentren en una escala unificada. Esto implica estandarizar las unidades y las distancias. En la figura 2 y la tabla 4 se observa la falta de estandarización de las escalas originales.

Tras la transformación logarítmica y el escalado robusto, los datos se distribuyeron como se observa en las figuras 3 y 4.

Figura 3.

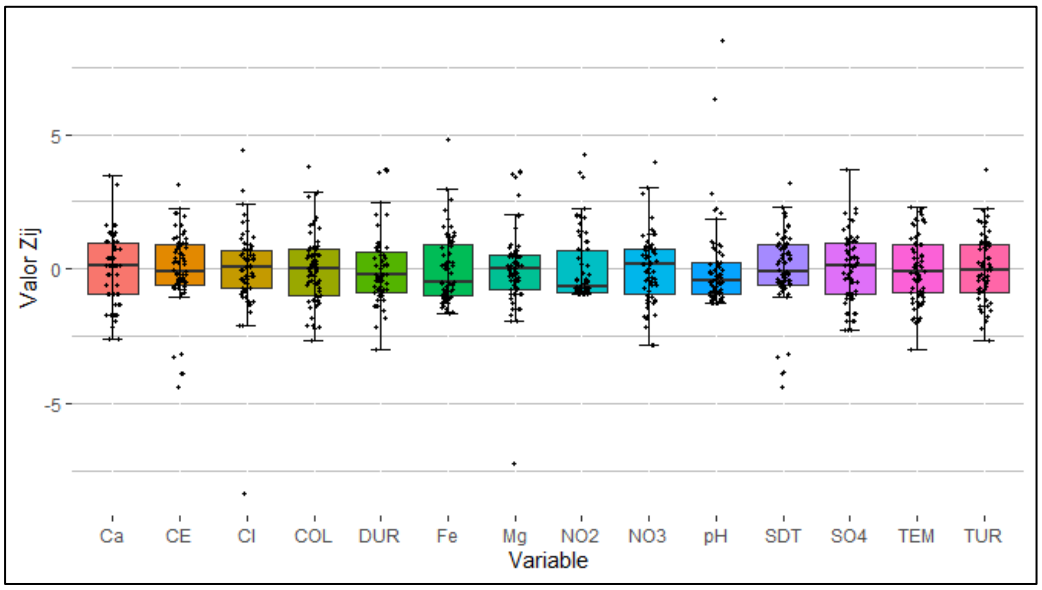
Histogramas de las variables estandarizadas



Nota. Datos distribuidos. Elaboración propia, realizado con R Studio.

Figura 4.

Gráficos de caja de las variables estandarizadas



Nota. Datos distribuidos. Elaboración propia, realizado con R Studio.

3.1.3. Análisis de componentes principales

El ACP se realizó con los datos estandarizados. La correlación, multicolinealidad, adecuación del muestreo y la interdependencia de variables se evaluaron previamente. Se redujo la dimensionalidad a siete componentes principales y siete variables influyentes.

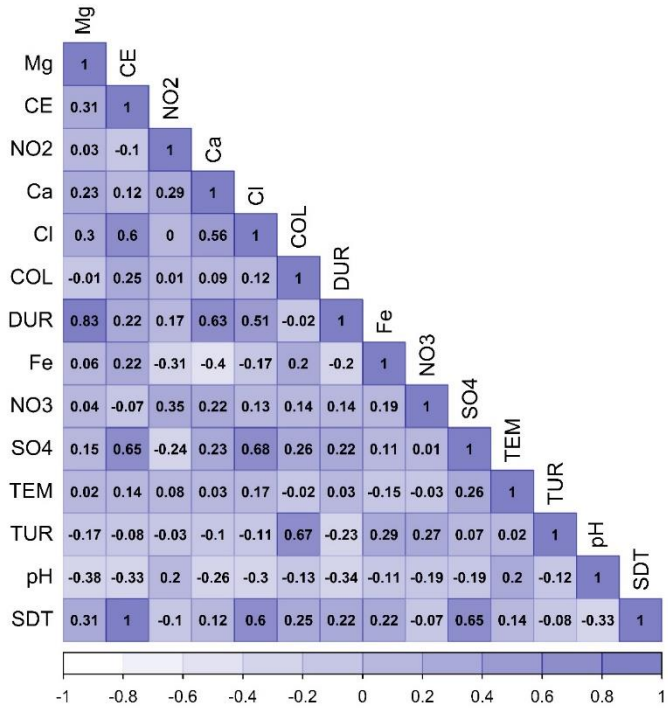
3.1.3.1. Correlaciones entre variables

En el correlograma de la figura 5 se puede observar la matriz triangular inferior de la matriz de correlaciones. En este se puede observar que se encontraron correlaciones fuertes (≥ 0.65) entre las variables CE – SDT (1.00), Mg – DUR (0.83), Cl – SO4 (0.68), COL – TUR (0.67), CE – SO4 (0.65) y SO4 – SDT (0.65), así como correlaciones moderadas (≥ 0.5) entre las variables Ca –

DUR (0.63), CE – Cl (0.6), CE – SDT (0.65) y DUR – Cl (0.51). Los demás pares de variables mostraban correlaciones débiles.

Figura 5.

Correlograma con correlación de Spearman



Nota. Matriz triangular inferior de la matriz de correlaciones. Elaboración propia, realizado con R Studio.

3.1.3.2. Multicolinealidad

Se evaluó la multicolinealidad entre las variables por medio del estadístico de Haitovsky, se obtuvo un $\chi_H^2 = 4.6 \times 10^{-6}$ y un valor p asociado de 2.9×10^{-314}. El valor cercano a cero de χ_H^2 indica la presencia de multicolinealidad, pero el valor p indica que no es severa para el análisis de ACP.

3.1.3.3. Adecuación del muestreo

La calidad de los datos para llevar a cabo el ACP y el AF se revisó con el índice de Kaiser-Meyer-Olkin (KMO). El índice obtenido fue de 0.604, con lo que se le valora como “mediocre” (0.60 – 0.69).

3.1.3.4. Independencia de las variables

La prueba de esfericidad de Bartlett fue utilizada para comprobar la hipótesis nula, H_0, de que los datos provengan de una población multivariada normal con variables independientes. Se obtuvo un estadístico de $\chi_B^2 = 991.18$, con un valor p asociado de 3.1×10^{-151}; con lo cual se descartó la hipótesis nula.

3.1.3.5. Componentes principales

Los valores propios, su varianza explicada y la varianza acumulada se muestran en la tabla 5. Para el criterio de corte compuesto por un valor propio mayor a 1, varianza explicada mayor a 0.1 y varianza acumulada explicada mayor a 0.60, se retienen cuatro CCPP; explicando un 69.50 % de la varianza total. Mientras que, para un criterio de corte $\lambda_0 = 0.7$, se retienen siete CCPP que explican un 88.78 % de la varianza total.

En la tabla 6 se observan los puntajes de los 14 componentes principales. En este aspecto, según el método B2 de descarte de variables propuesto por Jollife (1972), se consideran como variables influyentes: magnesio (Mg), conductividad eléctrica (CE), nitrito (NO2), hierro (Fe), nitrato (NO3), temperatura (TEM) y turbidez (TUR). En tanto que, según el método B4 de descarte de variables propuesto por Jollife (1972), se consideran como variables influyentes:

cloruro (Cl), hierro (Fe), nitrato (NO3), temperatura (TEM), nitrito (NO2), magnesio (Mg) y color (COL).

Tabla 5.

Valores propios y varianza explicada

	Valor propio	Varianza explicada	Varianza explicada acumulada
λ_1	3.9927	0.2852	0.2852
λ_2	2.3883	0.1706	0.4558
λ_3	1.8431	0.1317	0.5874
λ_4	1.5053	0.1075	0.6950
λ_5	0.9955	0.0711	0.7661
λ_6	0.9033	0.0645	0.8306
λ_7	0.8013	0.0572	0.8878
λ_8	0.5769	0.0412	0.9290
λ_9	0.3258	0.0233	0.9523
λ_{10}	0.2501	0.0179	0.9702
λ_{11}	0.2092	0.0149	0.9851
λ_{12}	0.1586	0.0113	0.9964
λ_{13}	0.0498	0.0036	1.0000
λ_{14}	0.0001	0.0000	1.0000

Nota. valores propios, su varianza explicada y la varianza acumulada. Elaboración propia, realizado con R Studio.

Tabla 6.

Puntajes de los componentes principales

	CP1	CP2	CP3	CP4	CP5	CP6	CP7
Mg	0.581	-0.317	0.043	-0.495	0.287	0.427	0.124
CE	0.819	0.349	-0.255	0.063	0.156	-0.185	0.221
NO2	-0.039	-0.489	0.408	0.365	0.417	-0.230	0.398
Ca	0.516	-0.563	0.316	0.181	-0.325	-0.097	-0.055
Cl	0.840	-0.131	-0.040	0.234	-0.160	-0.149	-0.162
COL	0.260	0.506	0.557	0.206	-0.208	0.233	0.364
DUR	0.650	-0.580	0.161	-0.270	0.057	0.286	0.027
Fe	0.035	0.684	0.113	-0.403	0.413	-0.035	-0.136
NO3	0.115	-0.068	0.714	0.022	0.372	-0.328	-0.404
SO4	0.735	0.311	-0.156	0.290	-0.097	-0.006	-0.235
TEM	0.164	-0.031	-0.191	0.620	0.324	0.491	-0.318
TUR	-0.059	0.551	0.679	0.157	-0.148	0.272	-0.003
pH	-0.513	-0.061	-0.251	0.511	0.262	0.092	0.174
SDT	0.820	0.351	-0.254	0.061	0.155	-0.183	0.219

	CP8	CP9	CP10	CP11	CP12	CP13	CP14
Mg	0.043	-0.124	-0.015	0.016	0.020	0.132	0.000
CE	-0.119	0.039	-0.048	0.049	0.104	-0.012	0.005
NO2	-0.118	-0.041	0.206	-0.090	-0.092	0.000	0.000
Ca	0.111	0.370	0.041	-0.020	0.103	0.066	0.000
Cl	0.129	-0.116	0.058	0.265	-0.208	0.014	0.000
COL	0.079	0.051	-0.240	-0.080	-0.145	0.000	0.000
DUR	0.166	-0.002	0.002	-0.011	0.031	-0.164	0.000
Fe	0.234	0.279	0.147	0.003	-0.106	-0.008	0.000
NO3	0.021	-0.102	-0.220	0.005	0.069	0.005	0.000
SO4	0.233	-0.180	0.122	-0.295	0.019	0.016	0.000
TEM	-0.284	0.143	-0.035	0.002	-0.041	-0.007	0.000
TUR	-0.061	-0.105	0.224	0.146	0.157	-0.008	0.000
pH	0.526	-0.014	-0.064	0.102	0.097	0.010	0.000
SDT	-0.116	0.037	-0.046	0.049	0.105	-0.014	-0.005

Nota. Se presentan los principales puntajes. Elaboración propia, realizado con R Studio.

3.2. Objetivo 2. Asignar pesos a las variables influyentes por medio de la aplicación de análisis factorial a los datos de las variables influyentes preseleccionadas con el análisis de componentes principales

Con las siete variables influyentes se realizó un análisis factorial (AF) y se procedió a asignar pesos a las variables y a los índices específicos derivados de los factores rotados por medio de una rotación *Varimax*.

3.2.1. Valores propios

Se realizó un segundo ACP con las siete variables retenidas. Se utilizó el mismo criterio de corte ($\lambda_0 = 0.7$), se retienen cuatro componentes principales; explicando un 69.50 % de la varianza total.

Tabla 7.

Valores propios y varianza explicada ACP - Influyentes

	Valor propio	Varianza explicada	Varianza explicada acumulada
$\boldsymbol{\lambda_1}$	1.5315	0.2852	0.2852
$\boldsymbol{\lambda_2}$	1.5140	0.1706	0.4558
$\boldsymbol{\lambda_3}$	1.3213	0.1317	0.5874
$\boldsymbol{\lambda_4}$	1.0258	0.1075	0.6950
$\boldsymbol{\lambda_5}$	0.6326	0.0711	0.7661
$\boldsymbol{\lambda_6}$	0.5654	0.0645	0.8306
$\boldsymbol{\lambda_7}$	0.4094	0.0572	0.8878

Nota. Se presentan los valores propios. Elaboración propia, realizado con R Studio.

3.2.2. Componentes rotados

El análisis factorial se llevó a cabo con el método de rotación conocido como *Varimax*, los componentes rotados (o factores) con su comunalidad (h), unicidad (u) e índice de complejidad se muestran en la tabla 8. El resumen asociado a la varianza explicada se muestra en la tabla 9.

Tabla 8.

Análisis de factores por rotación Varimax

	CR1	CR2	CR3	CR4	h	u	c
Mg	-0.112	0.147	0.818	-0.089	0.712	0.288	1.1
CE	0.173	-0.180	0.749	0.266	0.694	0.306	1.5
NO2	-0.306	0.828	-0.026	0.105	0.791	0.209	1.3
Fe	0.794	-0.167	0.261	-0.221	0.775	0.225	1.5
NO3	0.408	0.783	0.034	-0.083	0.787	0.213	1.5
TEM	-0.046	0.021	0.049	0.952	0.910	0.090	1.0
TUR	0.754	0.179	-0.312	0.161	0.723	0.277	1.6

Nota. Índice de complejidad. Elaboración propia, realizado con R Studio.

Tabla 9.

Varianza explicada por los componentes rotados

	CR1	CR2	CR3	CR4
Suma de cuadrados de puntajes	1.5033	1.4124	1.4003	1.0766
Varianza explicada	0.2148	0.2018	0.2000	0.1538
Proporción de varianza explicada	0.2788	0.2619	0.2597	0.1996

Nota. Resumen asociado a la varianza. Elaboración propia, realizado con R Studio.

3.2.3. Pesos temporales

Los puntajes de los componentes rotados se elevaron al cuadrado (tabla 10) y posteriormente se escalaron a la unidad (tabla 11). Los pesos temporales de las variables se obtuvieron seleccionando el mayor en cada renglón en la tabla 11.

Tabla 10.

Puntajes al cuadrado

	CR1	CR2	CR3	CR4
Mg	0.0125	0.0217	*0.6695*	0.0079
CE	0.0298	0.0324	*0.5610*	0.0708
NO2	0.0937	*0.6853*	0.0007	0.0110
Fe	*0.6306*	0.0278	0.0682	0.0487
NO3	0.1668	*0.6126*	0.0011	0.0069
TEM	0.0021	0.0004	0.0024	*0.9054*
TUR	*0.5678*	0.0322	0.0974	0.0258

Nota. Puntajes de los componentes rotados se elevaron al cuadrado. Elaboración propia, realizado con R Studio.

Tabla 11.

Puntajes escalados a la unidad

	CR1	CR2	CR3	CR4
Mg	0.0083	0.0154	*0.4781*	0.0073
CE	0.0198	0.0229	*0.4006*	0.0658
NO2	0.0623	*0.4852*	0.0005	0.0102
Fe	*0.4195*	0.0197	0.0487	0.0452
NO3	0.1110	*0.4337*	0.0008	0.0064
TEM	0.0014	0.0003	0.0017	*0.8410*
TUR	*0.3777*	0.0228	0.0695	0.0240

Nota. Puntajes escalados a la unidad. Elaboración propia, realizado con R Studio.

Tabla 12.

Pesos relativos

	Peso relativos (w_i)
Mg	0.1391
CE	0.1166
NO2	0.1412
Fe	0.1221
NO3	0.1262
TEM	0.2448
TUR	0.1099

Nota. Los pesos temporales se escalaron a la unidad; obteniéndose los pesos relativos. Elaboración propia, realizado con R Studio.

3.3. Objetivo 3. Transformar la escala de las variables seleccionadas, tomando como base los distintos "niveles" deseables que indica la literatura para las variables de estudio

La transformación de la escala de las variables Mg, CE, NO2, Fe, NO3 y TUR a una escala unitaria se llevó a cabo mediante un método similar al utilizado por (Flaieh, Mohammed-Ridha y Abdul-Ahad, 2014). La ecuación de transformación para las variables descritas se da a continuación:

$$q_i = \begin{cases} 1 - \frac{x_i}{P95_i - P5_i}, & yx_i \leq P95_i - P5_i \\ 0, & yx_i > P95_i - P5_i \end{cases} \quad (84)$$

A diferencia del método utilizado por los autores, el denominador en la ecuación es el rango entre los percentiles $P95$ y $P5$ de las medianas de todos los valores de referencia tomados en consideración para todos los usos contemplados para la variable dada x_i.

Tabla 13.

Valores máximos de tolerancia

	$P95 - P5$
Mg	324.75
CE	4955
NO2	9.826
Fe	3.41
NO3	17.325
TUR	203.4

Nota. Los rangos $P95 - P5$, para cada variable. Elaboración propia, realizado con R Studio.

La escala de la variable TEM se transformó por medio de una curva de aproximación y posterior ajuste con un modelo cuadrático. Para ello se tomó como valor de referencia la temperatura media de la aldea Balanya, según los datos disponibles públicamente del INE, y se asignaron los valores de calidad 1, 0.9, 0.7 y 0.3 para cero, una, dos y tres desviaciones con respecto a la temperatura media. Así, la ecuación obtenida del ajuste está dada por:

$$q_{TEM} = \begin{cases} -8.593396 + 1.165586 \times x_{TEM} - 0.035440 \times x_{TEM}^2, & 11.2 \leq x \leq 21.7 \\ 0, & x < 11.2 \ y \ x > 21.7 \end{cases} \quad (85)$$

3.4. Objetivo 4. Establecer la estructura del modelo y evaluar su confiabilidad mediante la comparación con otros índices de calidad aceptados mundialmente y criterios técnicos

Dos modelos de agregación fueron evaluados por comparación de correlación y comportamiento con otro par de índices de amplia trayectoria en quince observaciones viables.

La categorización del índice fue evaluada contra la categorización propuesta por Dinius mediante la prueba de correlación de Spearman.

3.4.1. Estructura matemática y su comparación

Dos propuestas de modelos de agregación fueron evaluadas, un modelo aditivo-ponderado y uno multiplicativo-ponderado. Ambos modelos fueron comparados coherentemente con los índices aditivo y multiplicativo ponderados propuestos por Dinius (1972, 1987), respectivamente, se utilizaron quince observaciones de la base de datos depurada.

Para el par de modelos de tipo multiplicativo se obtuvo un coeficiente de correlación de Pearson de 0.538 y para el par de modelos aditivos se obtuvo 0.523. Se comparó además una versión modificada sin coliformes del índice aditivo de Dinius (1972) y la correlación se mantuvo en 0.506 (ver figura 6).

Figura 6.

Valoración y categorización de la calidad

		O1	O2	O3	O4	O5	O6	O7	O8	O9	O10	O11	O12	O13	O14	O15
Dinius (1972)	**Valor**	46	43	44	46	39	39	31	41	37	44	33	35	35	35	43
	Categoría	4	4	4	4	5	5	5	4	5	4	5	5	5	5	4
Dinius (1972) modificado	**Valor**	68	63	66	68	58	59	46	60	54	66	49	52	51	52	64
	Categoría	2	2	2	2	3	3	5	2	3	2	5	3	3	3	2
Dinius (1987)	**Valor**	54	52	47	50	34	47	42	54	43	49	40	42	44	42	50
	Categoría	3	3	4	3	5	4	4	3	4	4	5	4	4	4	3
Modelo aditivo	**Valor**	81	70	71	84	60	49	56	61	48	75	61	64	73	64	52
	Categoría	1	1	1	1	2	3	2	2	3	1	2	2	2	2	3
Modelo multiplicativo	**Valor**	75	58	0	81	0	0	0	0	0	0	0	0	0	0	0
	Categoría	1	2	6	1	6	6	6	6	6	6	6	6	6	6	6

Nota. Resultados de la valorización y categorización. Elaboración propia, realizado con Microsoft Word.

Es importante notar que el modelo multiplicativo propuesto proporcionó valores distintos a cero para tres de los quince puntos evaluados, contra quince en el caso del modelo de Dinius en 1987. Por otro lado, los modelos aditivos proporcionaron valores distintos a cero para los quince puntos.

Se estableció la estructura aditiva para el modelo de agregación, a continuación, se muestra la ecuación propuesta:

$$ICA = 100 \times \sum_{i=7}^{7} w_i \times q_i \qquad (86)$$

3.4.2. Estructura categórica y su comparación

Con el uso de las 67 observaciones de la base de datos depurada, el método descrito por Ruiz-Picos, Kohlmann, Sedeño-Díaz y López-López (2017) y tomar en cuenta consideraciones adicionales para los límites de clase, se crearon categorías e intervalos de clase para la escala de cero a cien del modelo matemático, según la tabla 5.

Figura 7.

Categorías de calidad del agua

Intervalo	Categoría	Código de color
67 – 100	1—Excelente	
50 – 74	2—Buena, no afectada sensiblemente	
37 – 55	3—Regular	
25 – 41	4—Mala, contaminada	
12 – 27	5—Mala, muy contaminada	
0 – 14	6—Mala, extremadamente contaminada	

Nota. Cuando el valor del índice se encuentra dentro de dos categorías se asigna la categoría inferior. Elaboración propia, realizado con Microsoft Word.

La escala de categorización también se evaluó, se generaron valores con el modelo aditivo, el índice de Dinius (1972) y la versión modificada sin coliformes de este último. A las observaciones se les asignó las categorías correspondientes a sus valoraciones según la figura 7 y la categorización propuesta por Dinius para poder comparar las estructuras por medio de un coeficiente de correlación por rangos de Spearman (ρ_S).

Para la comparación del modelo aditivo contra el índice de este autor se obtuvo $\rho_S = 0.601$ y contra la versión modificada sin coliformes $\rho_S = 0.528$. Por su parte, para la comparación entre modelos multiplicativos se obtuvo $\rho_S = 0.638$. La Figura 6 muestra los valores utilizados para estas comparaciones.

3.5. Objetivo general. Proponer un modelo para estimar la calidad fisicoquímica del agua de los ríos Xayá y Pixcayá en Guatemala por medio de estadística multivariada

A través del ACP y de la técnica de descarte de variables propuesta por Jollife (1972) se dedujo que las variables influyentes en el conjunto de datos son magnesio (Mg), conductividad eléctrica (CE), nitrito (NO2), hierro (Fe), nitrato (NO3), temperatura (TEM) y turbiedad (TUR).

Al tomar en consideración únicamente las variables influyentes y al utilizar la técnica de AF con rotación *Varimax*, se obtuvieron los puntajes representativos de las variables en cada factor. Los pesos relativos de las variables se obtuvieron luego de una serie de transformaciones matemáticas de los puntajes asociados a las variables, siendo estos: 0.1391 – Mg, 0.1166 – CE, 0.1412 – NO2, 0.1221 – Fe, 0.1262 – NO3, 0.2448 – TEM y 0.1099 – TUR.

La transformación de escala de las siete variables se dividió en dos grupos. En el primer grupo se incluyó únicamente a la temperatura debido a su naturaleza. El ajuste con una función cuadrática para producir una parábola invertida permitió obtener la función de transformación de escala para la temperatura.

En el segundo grupo se incluyeron a todas las demás variables y la función de transformación se basó en una relación que se sustrae a la unidad (ver ecuación 84), en dicha relación el numerador es el valor que adquiere la variable y el denominador el rango definido por los percentiles 95 y 5 del conjunto de medianas obtenido de los valores de referencia por categoría.

La agregación de los subíndices y sus pesos se llevó a cabo por un modelo aditivo ponderado, el cual fue verificado mediante comparación con los índices propuestos por Dinius (1972, 1987). Así, el índice de calidad de agua se define de manera explícita por la ecuación 87 y se denomina como “ICAXP23”.

$$ICA = 100 \times \sum_{i=1}^{7} w_i \times q_i = \quad (87)$$
$$100 \times (0.1391 \times \left(1 - \frac{x_{Mg}}{324.75}\right) + 0.1166 \times \left(1 - \frac{x_{CE}}{4955}\right) + 0.1412 \times \left(1 - \frac{x_{NO2}}{9.826}\right) +$$
$$0.1221 \times \left(1 - \frac{x_{Fe}}{3.41}\right) + 0.1262 \times \left(1 - \frac{x_{NO3}}{17.325}\right) + 0.1099 \times \left(1 - \frac{x_{TUR}}{203.4}\right) +$$
$$0.2448 \times (-8.593396 + 1.165586 \times x_{TEM} - 0.035440 \times x_{TEM}^2)$$

4. DISCUSIÓN DE RESULTADOS

4.1. Análisis interno

La base de datos fue suministrada por una entidad privada que solicitó mantener anonimato. Dicha base de datos se recibió en físico y con una estructura de documentación de la información precaria, por lo que requirió su digitalización en una hoja de datos, en un formato lógico para la manipulación de la información.

Otro aspecto técnico no asociado a la base de datos, pero con relevancia técnica y científica fue la ausencia de declaración del método de análisis. Derivado de ello, a lo largo del análisis de datos se realizó una serie de inferencias con respecto a parámetros como los sólidos disueltos totales, el color y la turbidez.

Para estandarizar la base de datos, se realizó la suposición de que todos los parámetros (variables) y fechas estaban previstas para ser monitoreadas en los cinco puntos de muestreo. Así, la información quedó distribuida en 45 variables y 17 fechas asociadas a cinco puntos de muestro (los cuales no se encontraron georreferenciados).

De entrada, se observó una cantidad considerable de valores ausentes, falta de criterios técnicos en el registro de datos (límites, y cifras significativas); denotando un programa de monitoreo deficiente y, posiblemente, la utilización de distintos equipos y métodos no estandarizados.

La tasa de valores faltantes fue de 0.5373 sobre el total. Esta proporción, a toda vista inconcebible, provocó la toma de decisiones siguiente:

- Las variables con menos valores faltantes se retuvieron
- Las variables con hasta un 20 % más de valores faltantes con respecto al inciso anterior se retuvieron.

Los datos imputados según el segundo criterio corresponden a tres datos de magnesio (Mg) y nitrito (NO_2^-). Lo que en adelante generó la cantidad de observaciones ($n = 67$).

Posterior a estas imputaciones, se realizaron gráficos de cajas para evaluar cada variable, con los cuales se evidenció la existencia de datos atípicos. La imputación de estos datos atípicos se realizó con extrema precaución porque se sabe que inherentemente los indicadores de los ríos en Guatemala pueden variar notoriamente por las descargas y porque las técnicas estadísticas utilizadas buscan capturar la variabilidad del conjunto de datos.

Dicho esto, la imputación no se realizó únicamente al considerar los "valores extremos", sino que se consideraron los criterios científicos y la experiencia profesional.

Se encontró un valor inesperadamente bajo de dureza de $1.4\ \mathrm{mgCaCO_3/L}$, el cual no coincidía con los datos del punto de muestreo $97 \pm 32\ \mathrm{mgCaCO_3/L}$. El valor se verificó según una ecuación propuesta por APHA, AWWA y WEF (2017), la cual indica que la dureza puede ser estimada con las concentraciones de calcio, Ca, y magnesio, Mg, según la ecuación 88. Lo que generó una dureza calculada de $90 \pm 32.5\ \mathrm{mgCaCO_3/L}$ y un valor individual que solamente considera

la cantidad de calcio superaba el valor declarado de dureza; confirma la validez de la imputación por medio de la media.

$$DurCal = 2.497 * Ca + 4.118 * Mg \quad (88)$$

Otro valor atípico se encontró para el hierro, el cual era de 104mgFe/L, aunque improbable, no es imposible que provenga de factores antropogénicos en la cuenca. Sin embargo, el valor se encuentra unas cien veces por encima del intervalo del conjunto de datos ($0.74 \pm 0.573\text{mgFe/L}$), por lo que parecía tratarse de un problema de digitación y se imputó con la media.

El tercer caso se presentó en un dato de sulfato (SO_4^{2-}), con un valor de $374\ \text{mg}\ \text{SO}_4^{2-}\text{/L}$, excediendo el ya amplio intervalo del grupo dado por $53 \pm 33.6\ \text{mg}\ \text{SO}_4^{2-}\text{/L}$ en más de cinco desviaciones estándar. Las sales de sulfatos son en su mayoría insolubles en agua, no obstante, su presencia se puede atribuir a sales solubles de sodio, potasio y amonio (entre otras). Si las primeras dos sales estuvieren presentes, la cantidad de sodio o potasio hubiera incrementado sobre la media y, consecuentemente, la conductividad eléctrica también hubiera aumentado, pero no fue el caso. La presencia de sulfatos también produce la precipitación de iones calcio (II) y magnesio (II), pero nuevamente el valor de la dureza coincidente con la media dio pauta que no fue así; por lo que se imputó el dato.

El último caso se dio con un valor de temperatura de $678\ °C$, lo cual no es lógico puesto que el agua ebulle aproximadamente a $100°C$. Por ello, la discusión no se amplía y se imputa el valor con la media ($20.9\ °C$).

Con lo anterior, se logró obtener una base de datos estandarizada, libre de valores ausentes y lista para trabajar estadísticamente. Como de rutina, se dio

inicio al trabajo estadístico por medio de una exploración del conjunto de datos con estadística descriptiva. Los estadísticos descriptivos revelaron la falta de normalidad, elevada variabilidad, diferencia de unidades y de escalas en los datos.

Según Cuadras (2014), técnicas de ACP y AF son técnicas con un fundamento geométrico mediante el cual buscan similitudes o disimilitudes a través de conceptos de distancias. Además, Härdle y Simar (2003) indican que el ACP y el AF no son invariantes a la escala. Por ello, se deduce que las diferencias de escala y unidades favorecen unas variables sobre otras y detonan la necesidad de estandarizar la escala de las variables.

Dicho esto, para reducir el efecto de los valores extremos y las diferencias de los rangos "naturales" de cada variable, se aplicó la transformación logarítmica a cada variable, según la ecuación 89.

$$x_{if} = log_{10}(x_i + 1) \tag{89}$$

Nótese que a la transformación se le ha sumado la unidad, previo a aplicar el logaritmo, esto es porque el logaritmo de cero no está definido y, en caso de encontrarse un cero, se generan problemas en la transformación.

Cabe mencionar que previo a la transformación indicada se transformó, de manera independiente, la variable de potencial de hidronio (pH). La razón detrás de esto es que el pH es, en sí mismo, una transformación logarítmica (Skoog, 2014). Los detalles científicos se dejan para análisis del lector y basta con indicar que la transformación de la variable se da según la ecuación 90.

$$C_{H_3O^+} = 1 \times 10^{-pH} \tag{90}$$

La escala se estandarizó a través de la ecuación 91.

$$z_{if} = \frac{x_{if} - m_f}{s_f} \quad (91)$$

La funcionalidad de la transformación de las variables y la estandarización de las escalas se evidencia en la Figura 4, en la cual las cajas ya se encuentran aproximadamente centradas sobre una media cero de la escala z, mientras que retiene la dispersión característica de cada variable.

El ACP y el AF no tienen problema con la existencia de correlaciones entre variables. Sin embargo, es interesante explorar estas características para auxiliar la comprensión de los resultados que generen las técnicas (Härdle y Simar, 2003). Las correlaciones fueron evaluadas mediante la correlación de Spearman porque no existía una distribución normal de los datos y además se conocía la presencia de valores extremos.

En el correlograma presentado se observa la matriz triangular inferior de la matriz de correlaciones. Llamó la atención la correlación perfecta entre las variables CE y SDT, la cual se explica fácilmente. Los equipos de medición conocidos como “multiparamétricos”, ampliamente utilizados en campo, permiten la medición de una gran cantidad de parámetros, entre ellos la conductividad. Bruttel (2018), indica que los SDT pueden estimarse con facilidad a partir de la CE por medio de la ecuación 100, cuando se toma como referencia una disolución de cloruro de sodio 0.01 M.

$$SDT = 0.51 \times CE \quad (92)$$

Esto es evidentemente una idealización que no se cumple a cabalidad en muestras reales y, siendo la correlación perfecta, se dedujo que los SDT se determinaron de esta manera y que eran, por tanto, una variable redundante que había que seguir cercanamente por posible multicolinealidad.

Las correlaciones fuerte y moderada encontradas entre las variables Mg – DUR y Ca – DUR, respectivamente, se explica con la relación subyacente indicada anteriormente y representada en la ecuación 88. Aunque cabe resaltar que esta es nuevamente una idealización y no responde a todas las variables que contribuyen a la dureza.

Ahora bien, el color y la turbidez. Esta relación se puede explicar pensando en el color aparente, una forma de determinar color en la cual no se hace una filtración de la muestra para remover los sólidos suspendidos que provocan la turbidez. El color aparente recoge tanto información asociada al color real de una muestra, como a la turbidez de la muestra y dadas las condiciones cambiantes de un cuerpo lótico, no se espera una relación muy fuerte.

La multicolinealidad es de interés en ACP porque puede complicar la interpretación de los CCPP. Se evaluó la multicolinealidad a través de la prueba de Haitovsky y se obtuvo como estadístico 4.6×10^{-6}, con un p valor de 2.9×10^{-314}. El valor cercano a cero de χ_H^2 indica la presencia de multicolinealidad, pero el valor p indica que no es severa para el análisis de ACP y se presume que se puede ejecutar el ACP y el AF sin inconveniente. Además, las técnicas de factorización por regresión y las de rotación previenen que algunas variables con correlaciones elevadas puedan provocar dificultades.

La evaluación de adecuación del muestreo se realizó con la prueba de Kaiser-Meyer-Olkin (KMO), con el objetivo de revisar que proporción de la

varianza que puede ser atribuible a la varianza común. Se obtuvo un estadístico de 0.604. Lo cual supone una adecuación “mediocre” (0.60 – 0.69), aunque múltiples autores señalan 0.60 como un punto de partida para poder realizar el ACP.

Según este mismo autor la adecuación del muestreo parece estar relacionada al número de variables, observaciones y factores latentes y a la correlación general $\bar{r}$. Los efectos planteados por Kaiser, sin embargo, no tienen un parámetro de referencia y quedan sujetos por tanto a cada caso.

En términos del número de variables y observaciones, por cotidianeidad se sugiere contar con diez observaciones por variable, aunque autores como Comrey y Lee (1992) sugieren la siguiente escala de calidad del muestreo: 1000 – excelente, 500 – muy bueno, 300 – bueno, 100 – pobre y 50 – muy pobre.

Considerando que se contó con cerca de cinco observaciones por variable (un total de 67 observaciones para 14 variables); se dedujo que la cantidad de observaciones fue el factor que castigó el indicador.

La interdependencia de variables se evaluó con la prueba de esfericidad de Bartlett. Dicha prueba tiene como hipótesis nula que los datos provengan de una población multivariada normal con variables independientes, en dicho caso no habría varianza común y por tanto no tendría sentido realizar un ACP. El estadístico asociado fue de $\chi_B^2 = 991.18$, con un valor p asociado de 3.1×10^{-151}; con lo cual se descartó la hipótesis nula.

De los tres supuestos iniciales ausencia de multicolinealidad, adecuación del muestreo e interdependencia de las variables, solamente el segundo se encontró por debajo de lo deseable. Bajo estas condiciones se estableció que era

factible realizar el ACP y el AF, pero que los resultados derivados de estos podrían no ser concluyentes o en todo caso debían ser interpretados con especial cuidado.

Múltiples criterios han sido propuestos para poder tomar la decisión de cuántos componentes principales (CCPP) retener. Cuadras (2014) lista varios criterios, algunos de ellos versiones desagregadas del criterio compuesto propuesto por Tripathi y Singal (2019), que en todo caso provocaban el mismo efecto: retener pocas variables latentes. Sin embargo, Cuadras (2014) hace mención de que, aunque el criterio de Kaiser consiste en retener las componentes principales que cumplan con un valor propio mayor a la unidad ($\lambda_0 = 1$), estudios Monte Carlo han demostrado que el criterio de corte debería ser 0.7 ($\lambda_0 = 0.7$). Así, con dicho criterio de corte, se retuvieron siete componentes principales; explicando el 88.78 % de la varianza total.

Dicho valor de corte también sirvió para realizar el descarte de variables. Tripathi y Singal (2019) cita el método B4 de descartes de variables propuesto por Jollife en 1972. Sin embargo, al realizar este método se descartaron más variables que componentes principales y además se retuvo SDT (la cual se catalogó como redundante por ser una combinación lineal de la CE), por lo que, se decidió utilizar el criterio B2 de descarte de variables propuesto por Jollife (1972) para variables artificiales y ampliado en 1973 para variables reales.

El criterio de descarte consiste en utilizar los puntajes de las 14 variables en los 14 CCPP del ACP y asociar la variable con mayor puntaje en cada CCPP a dicho CCPP, realiza este proceso en orden creciente del valor propio ($\lambda_{14}, \lambda_{13}, \dots, \lambda_0$) hasta llegar al criterio de corte. Si la variable se repite, se utiliza el siguiente mayor puntaje. Las variables asociadas se descartan.

De tal manera, se retuvieron las variables magnesio, conductividad eléctrica, nitrito, hierro, nitrato, temperatura y turbidez (designadas como Mg, CE, NO2, Fe, NO3, TEM y TUR por practicidad).

Aunque no existe ningún criterio para saber si la selección fue o no adecuada, como auxilio a las precariedades de la adecuación del muestreo, se consideró como buen indicio la eliminación de SDT, por considerarse una variable redundante derivada de una combinación lineal de la CE. También se consideró como buen indicio que no se retuvieran conjuntamente TUR y COL, puesto que, como se indicó anteriormente, el color aparente contiene en sí mismo información asociada a la turbidez.

Las variables descartadas no fueron tomadas en consideración en las etapas posteriores. Los supuestos de multicolinealidad, adecuación del muestreo e interdependencia ya no se consideran para los siguientes ACP y AF puesto que ya ha sido manipulado el set de datos y se han cambiado sus características “naturales”.

La asignación de pesos inició con un segundo ACP que se llevó a cabo con la matriz de datos reducida. Utilizando nuevamente como criterio de corte $\lambda \geq 0.7$, se retuvieron cuatro componentes rotados (CCRR) que explicaron el 0.6950 de la varianza. Sin embargo, en este caso el interés era reducir la cantidad de variables con puntajes elevados en cada componente principal, por lo que se utilizó una rotación *Varimax*.

Los componentes rotados (CCRR) se muestran en la tabla 7. Se obtuvo un índice de complejidad de Hofmann de 1.4; lo que indicó que se requiere de uno a dos CCRR para explicar las variables latentes. Las variables que más se

alejaron de la unidad fueron CE, Fe, NO3 y TUR, aunque ninguna obtuvo puntajes mayores a 0.35 (negativo o positivo) en más de un factor.

Por otro lado, casi todas las comunalidades contaron con un valor superior a 0.7, con una comunalidad media de 0.7704, por lo que se concluyó que tanto las variables seleccionadas con el auxilio del primer ACP y del método “B2” propuesto por Jollife (1972), como la cantidad de factores retenidos en el segundo ACP explicaban de manera adecuada la variabilidad del conjunto de datos.

El proceso de asignación de pesos continuó según lo expuesto por Tripathi y Singal (2019), elevando al cuadrado los puntajes y escalándolos a la unidad (por medio de la división entre la suma de los cuadrados de los puntajes); los resultados asociados a este proceso se muestran en las tablas 10 y 11.

De los resultados anteriores se eligió, para cada variable, el mayor valor asociado a los factores (mayor valor del renglón de la tabla 11) y, el conjunto de valores conocidos como pesos temporales, se escalaron a la unidad, así los pesos relativos obtenidos para las variables fueron: Mg – 0.1391, CE – 0.1166, NO2 – 0.1412, Fe – 0.1221, NO3 – 0.1262, TEM – 0.2448 y TUR – 0.1099.

Al igual que todo el proceso de ACP y AF requirió de la unificación de escalas y culminó en una escala unitaria. Se propusieron transformaciones de escala para cada variable x_i. Estas nuevas escalas, convenientemente, también se planificaron unitarias.

Para este fin, se investigó una gran cantidad de literatura y normativas nacionales e internacionales, buscando valores de referencia, tolerancia o reglamentarios en siete usos del agua: consumo o uso humano, consumo

avícola, consumo bovino, consumo porcino, acuicultura y vida acuática, riego y aguas naturales o recreación.

La intención inicial de recopilar toda esa información era generar algún tipo de regresión para lograr un ajuste, sin embargo, se encontró que:

- no todas las variables seleccionadas eran de interés para todos los usos; provocando valores “faltantes”,
- para cada variable los valores asociados a las categorías normalmente presentaban concordancia,
- existían usos con criterios rigurosos y otros con criterios muy holgados y
- casi ningún uso declaraba valores de temperatura.

Por lo anterior se descartó la idea de regresión en todos los casos excepto para la temperatura.

La temperatura como indicador de calidad se describe en la mayoría de literatura en términos de una variación con respecto a un valor de referencia. Este valor, usualmente es la temperatura ambiente y alejarse de este en cualquier dirección produce una baja de calidad.

Ante esto, se tomó la temperatura media registrada en la Aldea Balanya de 2017 a 2020 según los datos publicados por el INE, así como su desviación estándar; siendo estas 16.4 °C y 1.5 °C. Al valor medio se le asignó una calidad de uno, y cada desviación estándar que se alejaba del valor medio se le asignó un decrecimiento geométrico: 1σ — 0.9, 2σ — 0.7 y 3σ — 0.3. Un ajuste con una ecuación cuadrática permitió obtener la parábola invertida: $q_{TEM} = -8.593396 + 1.165586 \times x_{TEM} - 0.035440 \times x_{TEM}^2$.

Para los demás datos, se concibió la calidad como una línea recta, la cual empezaría en el percentil 5 del rango asociado a la variable i y tendría un valor de uno, para terminar en el percentil 95 de dicho rango con un valor de cero. El rango se formuló al considerar, para cada variable, la mediana de cada uno de los siete usos y del conjunto de medianas se calcularon los percentiles 5 y 95 con el fin de evitar una sobreestimación del rango.

La estructura matemática que se utilizó para describir esa recta fue una relación simple que se sustraía a la unidad. La relación presenta en el denominador a la variable, x_i, y en el denominador al rango descrito anteriormente. Así, a cada variable se le atribuye una ecuación única.

Los pesos relativos y las variables selectas deben agregarse mediante algún modelo, los modelos aditivo-ponderado y multiplicativo-ponderado fueron seleccionados por su sencillez y capacidad de incorporación de los pesos o importancias atribuibles a las variables.

Las observaciones en la cuenca presentaban con mucha frecuencia una o dos variables que excedían la tolerancia y, en consecuencia, el modelo multiplicativo por su naturaleza penalizó con ceros muchas observaciones (ver figura 6), lo cual para un índice de uso específico (por ejemplo, un ICA para agua potable) resulta conveniente pues no debe quedar mucho margen de tolerancia. Además, también se consideró como contraproducente en el efecto concienciador que el índice podría llegar a tener en la sociedad; pues a cualquier persona le resultaría desalentador perder un examen y siempre con un cero.

El modelo aditivo, por su parte valoraba la calidad del agua con una mediana de 70 al aplicar el modelo en las 67 observaciones. Al evaluar los subíndices por separado se verificó que de manera general se podía indicar que

cinco de los siete parámetros cumplían con tolerancias incluso de agua potable y dado que el índice es de uso general la calificación se consideró válida.

Además, a partir de que las técnicas utilizadas en este estudio para la selección de variables y la asignación de pesos fueron el ACP y el AF que consisten combinaciones lineales de estructura matemática similar al modelo aditivo; se decidió utilizar un modelo aditivo-ponderado de agregación.

Seleccionar un modelo aditivo ponderado implica tener efectos de eclipsado (Abbasi y Abbasi, 2012). Pero se consideró que es deseable obtener una calificación a obtener ceros.

Para mitigar el efecto del eclipsado, se construyó una escala de categorización del agua. La escala se construyó con el método descrito por Ruiz-Picos, Kohlmann, Sedeño-Díaz y López-López (2017), se utilizaron las 67 observaciones. Los autores indican que el límite de clase entre la primera categoría (máxima) y la inferior es la mediana de las valoraciones del modelo, mientras que el límite de clase entre la segunda y la tercera está dado por el percentil 10 de las valoraciones del modelo; los límites de clase de las categorías cuatro, cinco y seis se obtienen al dividir entre cuatro el percentil 10.

Además, se consideró lo expuesto por Tercedor (1996) con respecto a la problemática de establecer límites estrictos de clase. Así, se decidió que los límites de clase definidos anteriormente tendrían una tolerancia del 5 %; permitiendo el traslape de categorías y definiendo que aquellas valoraciones que quedan entre dos categorías deben ser penalizadas asignándolas a la categoría más baja (ver figura 7).

Con lo anterior, se evidenció que las observaciones oscilaron entre las categorías “1” a la “3” y que la mediana de las observaciones se asignó a la categoría 2 (Buena, no afectada sensiblemente); lo cual en indicadores fisicoquímicos pareció ser correcto.

Merece la pena aclarar que las categorías no se relacionan con las de un histograma y por lo tanto sus intervalos, límites de clase e interpretación tampoco.

4.2. Validez externa

Antes de iniciar esta sección se debe tener en mente el principio de rigidez de los índices, el cual Abbasi y Abbasi (2012) indican que puede tomar lugar al intentar utilizar un índice desarrollado para una región, en otra región que cuenta con condiciones distintas y que en consecuencia requiere de otras variables que el índice no podrá acomodar.

Dicho lo anterior, se puede comparar de alguna manera el desempeño de las técnicas estadísticas utilizadas, más no se puede comparar el índice de manera ligera.

En esta investigación se utilizó una base de datos externa que contaba deficiencias severas y provocó el retiro de múltiples variables previo al análisis estadístico, entre ellas variables que múltiples autores utilizan dentro de sus índices de calidad del agua como la demanda química de oxígeno, la demanda bioquímica de oxígeno y los coliformes según lo expuesto por Calvo (2018) en varias secciones de su libro.

Previo a realizar cualquier manipulación de los datos para el ACP o el AF, Qian, *et al*., (2007) proponen realizar una transformación de los datos con una

transformación logarítmica y una estandarización (utilizando una ecuación similar a la normalización *z-score*) con el objetivo de reducir los efectos de escala. Este acercamiento fue utilizado en este estudio y su eficiencia se ha podido corroborar en la figura 4.

Coletti, *et al.*, (2010) utilizaron una matriz de correlaciones de Spearman para buscar relaciones subyacentes en los datos. Este acercamiento permitió deducir que existían variables provenientes de combinaciones lineales dentro del conjunto de datos y, más aún, permitieron utilizarlas como "cebo" en el ACP para inferir si el descarte de variables estaba llevándose de manera adecuada.

La adecuación del muestreo de 0.604 fue valorada como "mediocre" según el criterio expuesto por Kaiser (1974), aunque por rutina se acepta un 0.60 como punto de inicio. No obstante, al compararla con la obtenida por Shirodkar, *et.al.,* (2010) de 0.850 y por Yang, *et al*., (2014), de 0.79, se confirma que los resultados obtenidos del ACP y AF pueden verse comprometidos por la baja cantidad de observaciones.

Existen múltiples criterios para la retención de CCPP, algunos enfocados en el valor propio, otros en la varianza explicada, entre otros. En este estudio se contempló el uso del expuesto por Tripathi y Singal (2019), un criterio compuesto, que retiene los CCPP que cumplan con un valor propio mayor a la unidad, varianza explicada mayor a 0.10 y varianza explicada acumulada mayor a 0.60. Sin embargo, con este criterio compuesto se retuvieron únicamente cuatro CCPP, algo que es inusual para los índices de enfoque similar, los cuales se han compuesto históricamente por siete a catorce variables, además de retener únicamente el 69.50 % de la varianza total.

Por lo anterior, se probó con otros criterios, que incluyen los criterios desagregados de Tripathi y Singal (2019); se obtuvo la misma cantidad de CCPP. Finalmente, se utilizó el valor de corte expuesto por Cuadras (2014), con el que se retienen los CCPP que tengan un valor propio mayor 0.7; reteniendo siete CCPP.

Tripathi y Singal (2019) también citaron un método de descarte de variables, específicamente el método B4 propuesto por (Jolliffe,1974). Se realizó el descarte de variables con este método, pero se retuvieron variables identificadas como redundantes en el análisis de correlación (CE y COL), por lo que se ahondó en el trabajo de Jolliffe (1972) y se utilizó el método B2; logrando retener variables no redundantes.

Es importante hacer mención en este punto que los algoritmos de iteración para hallar los valores y vectores propios que utilizan los diferentes paquetes estadísticos provocan que los puntajes varíen en dirección; sin modificar las conclusiones. Para prevenir esto y hacer coincidir los valores de R con SPSS, se utilizaron los atributos específicos para la función *psych::principal()*, siendo estos: *rotate = "varimax", normalize = FALSE y eps = 1e-14*.

La disposición de los puntajes de las variables en los CCRR resultó interesante, el CR1 pareció asociarse a los sólidos suspendidos en el agua, el CR2 a nutrientes causantes de eutrofización, el CR3 a la carga iónica y el CR4, evidentemente, a la temperatura; por lo que se entiende una estructura bien diferenciada y coherente con la ciencia.

La transformación de escala de todas las variables, excepto la temperatura, es muy parecida a la transformación propuesta por (Flaieh, Mohammed-Ridha y Abdul-Ahad, 2014). No obstante, en esta investigación se

deseaba que la escala fuera proporcional a la calidad e inversamente proporcional a la polución, por lo que una ligera modificación se introdujo a la estructura.

Además, debido a que el índice propuesto por Flaieh, Mohammed-Ridha y Abdul-Ahad (2014) es específico y el propuesto en este estudio general, el concepto de valor estándar se cambió por un rango asociado a un conjunto de valores estándar. El comportamiento se mantiene lineal, pero la pendiente se ha modificado.

Los modelos de agregación aditivo-ponderado y multiplicativo-ponderado fueron evaluados. Ambos modelos se compararon contra los índices, aditivo y multiplicativo, propuestos por Dinius (1972,1987), tomó en consideración quince observaciones. La comparación entre los puntajes de los modelos multiplicativos mostró una correlación de Pearson de 0.538, mientras que para los modelos aditivos fue de 0.523.

Para los modelos también se compararon las categorías. Para dicho efecto, a las categorías propuestas por Dinius (1972,1987) se asignó una escala de Likert 1, 2, …, 6 para cada categoría de cada uso y luego se sacó la moda; la cual se consideró como la categoría “general”. Obteniéndose una correlación de Spearman de 0.601 entre modelos aditivos y de 0.638 entre modelos multiplicativos.

Además, se observó que el modelo valoró de forma consistente más bajo el agua que el propuesto en este estudio (ver figura 6), con medias de 39 para el índice y de 46 para el índice. Esta situación se infirió que provenía de la consideración de los coliformes fecales en los índices y para corroborarlo se utilizó una modificación de este índice sin los coliformes (obsérvese el principio

de rigidez), obteniéndose una media de 58 que fue más cercano a la media de 65 que entregó el índice propuesto. Aunque la valoración mejoró, los coeficientes de Pearson, para las valoraciones, y de Spearman para las categorías, no se modificaron de manera considerable; pasando de $r = 0.523$ a $r = 0.506$ y de $r_s = 0.601$ a $r_s = 0.528$.

En lo que respecta al principio de rigidez que expone los riesgos de la utilización de índices en lugares y condiciones ajenas a las que fueron creados, las correlaciones entre los modelos aditivos parecen ser normales e incluso alentadoras para considerar un buen desempeño del índice propuesto (Gil-Rodas, *et al*., 2023; Calvo-Brenes 2018).

CONCLUSIONES

1. La aplicación del ACP permitió la selección de siete variables influyentes de un total de catorce, siendo estas: magnesio, conductividad eléctrica, nitrito, hierro, nitrato, temperatura y turbidez.

2. El análisis factorial con rotación Varimax generó los coeficientes que se transformaron en los pesos de las variables, siendo los pesos 0.1391 para magnesio, 0.1166 para conductividad eléctrica, 0.1412 para nitrito, 0.1221 para hierro, 0.1262 para nitrato, 0.2448 para temperatura y 0.1099 para turbidez.

3. Las escalas de las variables se transformaron mediante dos técnicas, siendo la primera un ajuste cuadrático utilizado para la variable temperatura y la segunda una transformación por relación simple; permitiendo que todas las variables se expresen en términos de una escala unitaria.

4. Se estableció una estructura aditiva-ponderada para el modelo de agregación de las variables y se obtuvo un desempeño adecuado.

5. El índice de calidad del agua denominado “ICAXP23” se propuso como modelo para estimar la calidad del agua de las cuencas Xayá y Pixcayá de Guatemala.

6. La precariedad de la base de datos previno que este modelo contara con indicadores microbiológicos como los coliformes, los cuales juegan un papel importante en la valoración de la calidad del agua.

RECOMENDACIONES

1. Continuar, intensificar y mejorar el plan de muestreo de la institución, así como la documentación de datos de las variables de monitoreo de calidad del agua, para que el indicador de adecuación del muestreo de Kaiser-Meyer-Olkin alcance un valor superior a 0.80 y que la selección de parámetros sea más adecuada en próximos estudios que utilicen las técnicas de ACP y AF.

2. Comparar en futuros estudios diversas técnicas de rotación ortogonales en el AF, que permitan un mejor entendimiento de los factores y, consecuentemente, de la asignación de pesos a las variables.

3. Ampliar en futuros estudios los usos y la información de valores de referencia para que los rangos asociados a las transformaciones de las variables sean más certeros.

4. Complementar en futuros estudios el método de evaluación del desempeño del índice generado, para evitar la ambigüedad de la comparación con otros índices no fueron desarrollados para la cuenca o área geográfica de interés.

5. A la institución o el usuario del modelo ICAXP23, revisar periódicamente la validez del modelo y actualizarlo si es necesario.

6. A la institución o el usuario del modelo ICAXP23, monitorear el grupo coliforme de forma paralela a los parámetros del modelo y considerar su inclusión en una futura actualización del modelo.

REFERENCIAS

Abbasi, T., y Abbasi, S. (2012). *Water quality indices*. Elsevier Science.

Brown, R., McClelland, N., Deininger, R., y Landwehr, J. (1973). *Validating the WQI*. Scientific Research. 11 (23) 295-307

Brown, R., McClelland, N., Deininger, R., y Tozer, R. (1970). A water quality index: Do we dare? Water Sewage Works. Scientific Research .117 (17) 339-343.

Bruttel, P. (2018). *Conductometry – Conductivity Measurement*. Metrohm.

Calvo-Brenes, G. (2018). *Índices e indicadores sobre la calidad del agua* (1ª edición). Editorial Tecnológica de Costa Rica.

Coletti, C., Testezlaf, R., Ribeiro, T. A. P., Souza, R. T. G. D., y Pereira, D. D. A. (2010). Water quality index using multivariate factorial analysis. *Revista Brasileira de Engenharia Agrícola e Ambiental*, *14*(5), 517-522. https://doi.org/10.1590/S1415-43662010000500009

Comrey, A. L., y Lee, H. B. (1992). *A first course in factor analysis* (2ª edición). Psychology Press.

Cuadras, C. (2014). *Nuevos métodos de análisis multivariante*. CMC Editions. https://gc.scalahed.com/recursos/files/r161r/w24899w/Semana5/METODOS_S5.pdf

Dinius, S. H. (1972). Social accounting system for evaluating water resources. *Water Resources Research*, *8*(5), 1159-1177. https://doi.org/10.1029/WR008i005p01159

Dinius, S. H. (1987). Design of an index of water quality. *Journal of the American Water Resources Association*, *23*(5), 833-843. https://doi.org/10.1111/j.1752-1688.1987.tb02959.x

Flaieh, H., Mohammed-Ridha, M., y Abdul-Ahad, M. (2014). Assessing Tigris River Water Quality in Baghdad City Using Water Quality Index and Multivariate Statistical Analysis. *International Journal of Engineering Sciences y Research Technology*, *3*(7), 687-699.

Gil-Rodas, N., Calvo-Brenes, G., Guerra, A., y Perdomo, A. (2021). Water quality assessment of six rivers of the Pacific side of Guatemala. *Environmental Earth Sciences*, *80*(5), 196. https://doi.org/10.1007/s12665-021-09505-w

Grossman, S. I. (2012). *Algebra lineal* (7ª edición). McGraw-Hill.

Härdle, W., y Simar, L. (2015). *Applied multivariate statistical analysis* (5ª edición). Springer.

Jolliffe, I. T. (1972). Discarding Variables in a Principal Component Analysis. I: Artificial Data. *Applied Statistics*, *21*(2), 160. https://doi.org/10.2307/2346488

Jolliffe, I. T. (1974). Discarding Variables in a Principal Component Analysis. II: Real Data. *Applied Statistics*, *22*(1), 21. https://doi.org/10.2307/2346300

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, *35*(4), 401-415. https://doi.org/10.1007/BF02291817

Kaiser, H. F., y Rice, J. (1974). Little Jiffy, Mark Iv. *Educational and Psychological Measurement*, *34*(1), 111-117. https://doi.org/10.1177/001316447403400115

Landwehr, J. (1974). *Water Quality Indices—Construction and Analysis*. University of Michigan.

Ley de derechos ciudadanos a la salud. Decreto 90-97. (7 de noviembre de 1997). Código de Salud. Diario de Centroamérica. No. 71 tomo CCLVII.

Ley de Creación del Ministerio de Ambiente y Recursos Naturales. Decreto 90-2000. (11 de diciembre de 2000). Congreso de la República de Guatemala. Diario de Centroamérica No. 54 tomo CCLV.

Ley de Protección y Mejoramiento del Medio Ambiente. Decreto 68-86. (19 de diciembre de 1986). Congreso de la República de Guatemala. Diario de Centroamérica No. 27 tomo CCXXX.

Qian, Y., Migliaccio, K. W., Wan, Y., y Li, Y. (2007). Surface water quality evaluation using multivariate methods and a new water quality index in the Indian River Lagoon, Florida: surface water quality evaluation using multivariate methods. *Water Resources Research*, *43*(8). 124-176. https://doi.org/10.1029/2006WR005716

Ruiz-Picos, R. A., Kohlmann, B., Sedeño-Díaz, J. E., y López-López, E. (2017). Assessing ecological impairments in Neotropical rivers of Mexico: Calibration and validation of the Biomonitoring Working Party Index.

International Journal of Environmental Science and Technology, *14*(9), 1835-1852. https://doi.org/10.1007/s13762-017-1299-x

Sharipov, R. A. (1996). *Course of Lineal Algebra and Multidimensional Geometry*. Bashkir State University.

Shirodkar, P. V., Pradhan, U. K., Fernandes, D., Haldankar, S., y Rao, G. S. (2010). Influence of anthropogenic activities on the existing environmental conditions of Kandla Creek (Gulf of Kutch). *Current Science*, *80*(6), 815-828.

Spiegel, M. (2009). *Estadística*. McGraw-HILL.

Triola, M. F. (2018). *Estadística* (12ª edición). Pearson Education.

Tripathi, M., y Singal, S. K. (2019). Allocation of weights using factor analysis for development of a novel water quality index. *Ecotoxicology and Environmental Safety*, *183*, 109-510. https://doi.org/10.1016/j.ecoenv.2019.109510.

Walpole, R. E., Myers, R. H., y Myers, S. L. (2012). *Probabilidad y estadística para ingeniería y ciencias* (9ª edición). Pearson Educación.

Walski, T., y Parker, Frank. (1974). Consumers Water Quality Index. *Journal of the Environmental Engineering Division*, 100(3), 593-611.

Yang, W., Zhao, Y., Wang, D., Wu, H., Lin, A., y He, L. (2020). Using Principal Components Analysis and IDW Interpolation to Determine Spatial and Temporal Changes of Surface Water Quality of Xin'anjiang River in

Huangshan, China. *International Journal of Environmental Research and Public Health*, 17(8), 2942. https://doi.org/10.3390/ijerph17082942

APÉNDICES

Apéndice 1.

Valores de referencia para magnesio

Uso	Valor	Unidades	Fuente
Consumo/uso humano	70	mg/L	OMS (2021)
Consumo/uso humano	100	mg/L	COGUANOR (2013)
Consumo/uso humano	30	mg/L	Indian Standards Institute (1982)
Riego	61	mg/L	Ayres y Westcot (1985)
Consumo aves de corral	250	mg/L	Ayres y Westcot (1985)
Consumo aves de corral	125	mg/L	Hess y Macklin (2019)
Consumo aves de corral	125	mg/L	Fairchild y Ritz (2015)
Consumo porcino	250	mg/L	Ayres y Westcot (1985)
Consumo bovino	400	mg/L	Ayres y Westcot (1985)
Aguas naturales y recreación	120	mg/L	Boyd y Tucker (1998)
Aguas naturales y recreación	30	mg/L	Summerfelt (1998)
Aguas naturales y recreación	4	mg/L	Bridgewater, et al. (2017)

Nota. Se presentan los valores de referencia para magnesio. Elaboración propia, realizado con Microsoft Excel.

Apéndice 2.

Valores de referencia para conductividad eléctrica

Uso	Valor	Unidades	Fuente
Consumo/uso humano	2500	µS/cm	OMS (2021)
Consumo/uso humano	1500	µS/cm	COGUANOR (2013)
Consumo/uso humano	1500	µS/cm	Ministerio de Salud de Perú (2011)
Riego	700	µS/cm	Jeong, H., Kim, H., y Jang, T. (2016)
Riego	1400	µS/cm	Jeong, H., Kim, H., y Jang, T. (2016)
Riego	3000	µS/cm	Jeong, H., Kim, H., y Jang, T. (2016)
Riego	1000	µS/cm	Jeong, H., Kim, H., y Jang, T. (2016)
Riego	3000	µS/cm	Jeong, H., Kim, H., y Jang, T. (2016)
Riego	2250	µS/cm	Indian Standards Institute (1982)
Riego	3000	µS/cm	Presidencia de la República de Ecuador (2015)
Riego	1500	µS/cm	Instituto Nacional de Normalización de Chile (1987)
Consumo aves de corral	5000	µS/cm	Ayres y Westcot (1985)
Consumo aves de corral	6000	µS/cm	Hess y Macklin (2019)
Consumo aves de corral	6000	µS/cm	Meehan, Stockam y Mostrom (2021)
Consumo porcino	8000	µS/cm	Ayres y Westcot (1985)
Consumo porcino	6000	µS/cm	Pfost y Fulhage (2001)
Consumo porcino	5000	µS/cm	Meehan, Stockam y Mostrom (2021)
Consumo bovino	8000	µS/cm	Ayres y Westcot (1985)
Consumo bovino	6000	µS/cm	Pfost y Fulhage (2001)
Consumo bovino	5000	µS/cm	Meehan, Stockam y Mostrom (2021)
Acuicultura y vida acuática	5000	µS/cm	Presidencia de la República de El Salvador (1987)
Acuicultura y vida acuática	1000	µS/cm	Indian Standards Institute (1982)
Aguas naturales y recreación	200	µS/cm	Summerfelt (1998)
Aguas naturales y recreación	1500	µS/cm	Bridgewater, et al. (2017)

Nota. Se presentan los valores considerados. Elaboración propia, realizado con Microsoft Excel.

Apéndice 3.

Valores de referencia para nitrito

Uso	Valor	Unidades	Fuente
Consumo/uso humano	0.2	mg N/L	Jeong, H., Kim, H., y Jang, T. (2016)
Consumo/uso humano	0.9	mg N/L	COGUANOR (2013)
Consumo/uso humano	1	mg N/L	Arizona Department of State (2016)
Consumo/uso humano	0.06	mg N/L	OECD (2011)
Consumo/uso humano	0.01	mg N/L	Oficina Nacional de Normalización de la República de Cuba (2014)
Consumo/uso humano	1	mg N/L	Presidencia de la República de Ecuador (2015)
Consumo/uso humano	3.2	mg N/L	Manitoba (2011)
Consumo/uso humano	0.9	mg N/L	Ministerio de Salud de la República de Perú (2011)
Riego	0.3	mg N/L	OECD (2011)
Consumo aves de corral	10	mg N/L	Mukhtar (1998)
Consumo aves de corral	4	mg N/L	Hess y Macklin (2019)
Consumo aves de corral	10	mg N/L	Pfost y Fulhage (2001)
Consumo aves de corral	1	mg N/L	Presidencia de la República de Ecuador (2015)
Consumo aves de corral	10	mg N/L	Manitoba (2011)
Consumo porcino	10	mg N/L	Ayres y Westcot (1985)
Consumo porcino	10	mg N/L	Mukhtar (1998)
Consumo porcino	10	mg N/L	Pfost y Fulhage (2001)
Consumo porcino	1	mg N/L	Ecuador (2015)
Consumo porcino	10	mg N/L	Meehan, Stockam y Mostrom (2021)
Consumo porcino	10	mg N/L	Manitoba Water Stewardship (2011)
Consumo bovino	10	mg N/L	Ayres y Westcot (1985)
Consumo bovino	10	mg N/L	Mukhtar (1998)
Consumo bovino	10	mg N/L	Pfost y Fulhage (2001)
Consumo bovino	10	mg N/L	Meehan, Stockam y Mostrom (2021)
Consumo bovino	1	mg N/L	Presidencia de la República de Ecuador (2015)
Consumo bovino	10	mg N/L	Manitoba (2011)
Acuicultura y vida acuática	3.5	mg N/L	Boyd y Tucker (1998)
Acuicultura y vida acuática	0.5	mg N/L	Buttner, Soderberg y Terlizzi (1993)
Acuicultura y vida acuática	4.5	mg N/L	Lazur (2007)
Acuicultura y vida acuática	0.01	mg N/L	OECD (2011)
Acuicultura y vida acuática	0.06	mg N/L	Manitoba (2011)
Aguas naturales y recreación	0.12	mg N/L	OECD (2011)

Nota. Se presentan los valores considerados. Elaboración propia, realizado con Microsoft Excel.

Apéndice 4.

Valores de referencia para hierro

Uso	Valor	Unidades	Fuente
Consumo/uso humano	0.3	mg/L	OMS (2021)
Consumo/uso humano	0.3	mg/L	COGUANOR (2013)
Consumo/uso humano	0.3	mg/L	Indian Standards Institute (1982)
Consumo/uso humano	1	mg/L	OECD (2011)
Consumo/uso humano	0.3	mg/L	Instituto Nacional de Normalización de Chile (2014)
Consumo/uso humano	0.3	mg/L	Presidencia de la República de Ecuador (2015)
Consumo/uso humano	0.3	mg/L	Manitoba (2011)
Consumo/uso humano	0.3	mg/L	Ministerio de Salud de la República de Perú(2011)
Riego	5	mg/L	Jeong, Kim y Jang (2016)
Riego	5	mg/L	Jeong, Kim y Jang (2016)
Riego	5	mg/L	Jeong, Kim y Jang (2016)
Riego	3	mg/L	Jeong, Kim y Jang (2016)
Riego	2	mg/L	Jeong, Kim y Jang (2016)
Riego	2	mg/L	Jeong, Kim y Jang (2016)
Riego	5	mg/L	Ayres y Westcot (1985)
Riego	5	mg/L	Pescod (1992)
Riego	5	mg/L	OECD (2011)
Riego	5	mg/L	Presidencia de la República de Ecuador (2015)
Riego	5	mg/L	Manitoba (2011)
Riego	5	mg/L	Instituto Nacional de Normalización de Chile (1987)
Riego	0.3	mg/L	Ministerio de Medio Ambiente y Recursos Naturales (2012)

Continuación del apéndice 4.

Uso	Valor	Unidades	Fuente
Consumo aves de corral	0.3	mg/L	Hess y Macklin (2019)
Consumo aves de corral	0.3	mg/L	Fairchild y Ritz (2015)
Consumo aves de corral	0.3	mg/L	Ministerio de Medio Ambiente y Recursos Naturales (2012)
Consumo bovino	1	mg/L	Presidencia de la República de Ecuador (2015)
Consumo bovino	0.3	mg/L	Ministerio de Medio Ambiente y Recursos Naturales (2012)
Consumo porcino	1	mg/L	Ecuador (2015)
Consumo porcino	0.3	mg/L	Ministerio de Medio Ambiente y Recursos Naturales (2012)
Consumo aves de corral	1	mg/L	Ecuador (2015)
Acuicultura y vida acuática	1	mg/L	Boyd y Tucker (1998)
Acuicultura y vida acuática	1	mg/L	OECD (2011)
Acuicultura y vida acuática	0.3	mg/L	Ecuador (2015)
Acuicultura y vida acuática	0.3	mg/L	Manitoba Water Stewardship (2011)
Acuicultura y vida acuática	0.3	mg/L	Ministerio de Medio Ambiente y Recursos Naturales (2012)
Aguas naturales y recreación	0.2	mg/L	Boyd y Tucker (1998)
Aguas naturales y recreación	0.7	mg/L	Boyd (2015)
Aguas naturales y recreación	3	mg/L	OECD (2011)

Nota. Se presentan los valores considerados. Elaboración propia, realizado con Microsoft Excel.

Apéndice 5.

Valores de referencia para nitrato

Uso	Valor	Unidades	Fuente
Consumo/uso humano	11	mg N/L	OMS (2021)
Consumo/uso humano	11	mg N/L	COGUANOR (2013)
Consumo/uso humano	10	mg N/L	Indian Standards Institute (1982)
Consumo/uso humano	10	mg N/L	Administrative Rules Division(2016)
Consumo/uso humano	4	mg N/L	OECD (2011)
Consumo/uso humano	10	mg N/L	Oficina Nacional de Normalización de la República de Cuba (2014)
Consumo/uso humano	10	mg N/L	Presidencia de la República de Ecuador (2015)
Consumo/uso humano	10	mg N/L	Manitoba (2011)
Consumo/uso humano	11	mg N/L	Ministerio de Salud de la República de Perú (2011)
Riego	10	mg N/L	Ayres y Westcot (1985)
Riego	20	mg N/L	OECD (2011)
Consumo aves de corral	23	mg N/L	Mukhtar (1998)
Consumo aves de corral	25	mg N/L	Hess y Macklin (2019)
Consumo aves de corral	25	mg N/L	Fairchild y Ritz (2015)
Consumo aves de corral	23	mg N/L	Pfost y Fulhage (2001)
Consumo aves de corral	10	mg N/L	Presidencia de la República de Ecuador (2015)
Consumo porcino	23	mg N/L	Ayres y Westcot (1985)
Consumo porcino	23	mg N/L	Mukhtar (1998)
Consumo porcino	23	mg N/L	Pfost y Fulhage (2001)
Consumo porcino	23	mg N/L	Meehan, Stockam y Mostrom (2021)
Consumo porcino	10	mg N/L	Presidencia de la República de Ecuador (2015)
Consumo bovino	23	mg N/L	Ayres y Westcot (1985)
Consumo bovino	23	mg N/L	Mukhtar (1998)
Consumo bovino	23	mg N/L	Pfost y Fulhage (2001)
Consumo bovino	23	mg N/L	Meehan, Stockam y Mostrom (2021)
Consumo bovino	10	mg N/L	Presidencia de la República de Ecuador (2015)
Acuicultura y vida acuática	1.5	mg N/L	OECD (2011)
Acuicultura y vida acuática	13	mg N/L	Manitoba (2011)
Aguas naturales y recreación	2	mg N/L	Boyd y Tucker (1998)
Aguas naturales y recreación	8	mg N/L	OECD (2011)

Nota. Se presentan los valores considerados. Elaboración propia, realizado con Microsoft Excel.

Apéndice 6.

Valores de referencia para turbiedad

Uso	Valor	Unidades	Fuente
Consumo/uso humano	5	UNT	OMS (2021)
Consumo/uso humano	15	UNT	COGUANOR (2013)
Consumo/uso humano	5	UNT	Indian Standards Institute (1982)
Consumo/uso humano	1	UNT	Administrative Rules Division (2016)
Consumo/uso humano	10	UNT	Oficina Nacional de Normalización de la República de Cuba (2014)
Consumo/uso humano	10	UNT	Presidencia de la República de Ecuador (2015)
Consumo/uso humano	1	UNT	Manitoba (2011)
Consumo/uso humano	5	UNT	Ministerio de Salud de Perú (2011)
Riego	2	UNT	Jeong, Kim y Jang (2016)
Riego	2	UNT	Jeong, Kim y Jang (2016)
Riego	2	UNT	Jeong, Kim y Jang (2016)
Riego	10	UNT	Jeong, Kim y Jang (2016)
Riego	10	UNT	Ayres y Westcot (1985)
Consumo porcino	30	UNT	Pfost y Fulhage (2001)
Consumo bovino	30	UNT	Pfost y Fulhage (2001)
Aguas naturales y recreación	250	UNT	Presidencia de la República de El Salvador (1987)

Nota. Se presentan los valores considerados. Elaboración propia, realizado con Microsoft Excel.

Apéndice 7.

Matriz de coherencia

Elementos	Problema Estadístico	Pregunta de investigación	Objetivos	Metodología	Resultados	Conclusiones	Recomendaciones
Específicos	Se desconoce qué parámetros de calidad de agua tienen mayor influencia para estimar la calidad de agua de los ríos Xayá y Pixcayá de Guatemala.	¿Cuáles son las variables (parámetros) más influyentes dentro del modelo para la determinación de un índice de calidad fisicoquímico del agua?	Seleccionar las variables influyentes mediante la reducción de la dimensionalidad, con las técnicas de análisis de componentes principales.	Análisis de componentes principales.	Siete variables de un conjunto de catorce fueron seleccionadas utilizando la técnica ACP. Las variables seleccionadas fueron magnesio, conductividad eléctrica, nitrito, hierro, nitrato, temperatura y turbidez.	Con el auxilio de la técnica de ACP siete variables influyentes de un total de catorce fueron seleccionadas, siendo estas magnesio, conductividad eléctrica, nitrito, hierro, nitrato, temperatura y turbidez.	Continuar, intensificar y mejorar el plan de muestreo y documentación de datos de las variables de monitoreo de calidad del agua, para que el indicador de adecuación del muestreo de Kaiser-Meyer-Olkin alcance un valor superior a 0.80 y que la selección de parámetros sea más adecuada en próximos estudios que utilicen las técnicas de ACP y AF.
	Se desconoce cuál es la proporción relativa en la que impacta cada parámetro influyente para estimar la calidad del agua de los ríos Xayá y Pixcayá de Guatemala.	¿Cuál es la importancia relativa de cada variable?	Asignar pesos a las variables influyentes por medio de la aplicación de análisis factorial a los datos de las variables influyentes preseleccionadas con el análisis de componentes principales.	Análisis factorial con rotación *Varimax*	Se asignaron los pesos 0.1445 para magnesio, 0.1211 para conductividad eléctrica, 0.1479 para nitrito, 0.1361 para hierro, 0.1322 para nitrato, 0.1955 para temperatura y 0.1226 para turbidez por medio de la transformación unitaria de los puntajes obtenidos del AF.	La técnica de análisis factorial sirvió para la asignación de los pesos a las variables, siendo estos 0.1445 para magnesio, 0.1211 para conductividad eléctrica, 0.1479 para nitrito, 0.1361 para hierro, 0.1322 para nitrato, 0.1955 para temperatura y 0.1226 para turbidez.	Utilizar y comparar diversas técnicas de rotación ortogonales en el AF que permitan un mejor entendimiento de los factores y, consecuentemente, de la asignación de pesos a las variables.

Continuación de apéndice 7.

Elementos	Problema Estadístico	Pregunta de investigación	Objetivos	Metodología	Resultados	Conclusiones	Recomendaciones
Específicos	Se desconocen los anamorfismos matemáticos de los parámetros asociados que permiten estimar la calidad del agua de los ríos Xayá y Pixcayá de Guatemala.	¿Qué anamorfismos de las variables (parámetros) permiten una transformación de escala que se alineen con las bases científicas de la química y la microbiología?	Transformar la escala de las variables seleccionadas, tomando como base los distintos "niveles" deseables que indica la literatura para las variables de estudio.	Regresión lineal simple, otras técnicas de transformación	Debido a la multiplicidad de usos que se le pueden dar al agua que se obtiene de las cuencas Xayá - Pixcayá y a la dificultad de valorar cada uno, no se encontró la regresión como una técnica adecuada para la transformación de escala de la mayoría de las variables. La única variable que presentó un comportamiento modelable por medio de regresión fue la temperatura; para la cual se generó una parábola invertida. La escala de las demás variables se transformó a través de ecuaciones matemáticas sencillas y un amplio bagaje de literatura científica.	Las escalas de las variables se transformaron mediante dos técnicas, siendo la primera un ajuste cuadrático utilizado para la variable temperatura y la segunda una transformación por relación simple; permitiendo que todas las variables se expresen en términos de una escala unitaria.	Ampliar los usos y la información de valores de referencia para que en futuros estudios los rangos asociados a las transformaciones de las variables sean más certeros.
	Se desconoce el modelo que permite estimar la calidad del agua de los ríos Xayá y Pixcayá de Guatemala.	¿Cuál es la estructura del modelo que permite estimar la calidad fisicoquímica del agua en los ríos Xayá y Pixcayá de Guatemala a través de estadística?	Establecer la estructura del modelo y evaluar su confiabilidad mediante la comparación con otros índices de calidad aceptados mundialmente y criterios técnicos.	Comparación de correlación por pares utilizando el coeficiente de correlación de Pearson y el coeficiente de correlación de Spearman	La estructura multiplicativa se comparó con el índice de Dinius (1987) y se obtuvo una correlación de 0.539, pero se encontró muy rigurosa. La estructura aditiva se comparó con el índice de Dinius (1972) y se obtuvo una correlación de 0.496. La ambigüedad inherente se mitigó con una escala, los rangos se compararon y se obtuvo una correlación de Spearman de 0.625. La estructura aditiva coincide con la estructura de ACP y AF.	Se estableció una estructura aditiva-ponderada para el modelo de agregación de las variables y se obtuvo un desempeño adecuado.	Considerar otros métodos de evaluación del desempeño del índice propuesto, para no introducir sesgo por comparación con otros índices no fueron desarrollados para la cuenca o área geográfica de interés.

Continuación de apéndice 7.

Elementos	Problema Estadístico	Pregunta de investigación	Objetivos	Metodología	Resultados	Conclusiones	Recomendaciones
General	No se cuenta con un modelo probabilístico para estimar la calidad de agua de ríos de Guatemala.	¿Es posible proponer un modelo que permita estimar el grado de calidad del agua en los ríos Xayá y Pixcayá en Guatemala con base en los distintos parámetros de calidad del agua (variables regresoras)?	Proponer un modelo generado a través de análisis multivariado para estimar la calidad del agua de los ríos Xayá y Pixcayá de Guatemala.	ACP, AF, regresión lineal, comparación por pares (Pearson y Spearman)	Se propuso el modelo $ICAXP23 = 100\sum_{i=0}^{7} w_i \times q_i$ Donde: w_i: peso relativo de cada variable q_i: valor transformado de cada variable. Y una escala complementaria de valoración – categorización de la calidad del agua. Encontrando su desempeño coherente con la literatura científica y las condiciones fisicoquímicas de la cuenca.	El índice de calidad del agua denominado ICAXP23 se propuso como modelo para estimar la calidad del agua de las cuencas Xayá y Pixcayá de Guatemala.	Revisar periódicamente la validez del modelo ICAXP23 y actualizarlo si es necesario. Monitorear el grupo coliforme de forma paralela a los parámetros del ICAXP23 y considerar su inclusión en una futura actualización del modelo.

Nota. Se presentan los valores considerados. Elaboración propia, realizado con Microsoft Excel.