Vol. 6 (1) pp. 54-63, enero-junio 2016

REVIEW ARTICLE

Revisión de los métodos estadísticos multivariados usados en el análisis de calidad de aguas

Ingry Natalia Gómez Miranda[1]*, Gustavo Antonio Peñuela Mesa[1]

[1]Grupo de Diagnóstico y Control de la Contaminación - GDCON, Escuela Ambiental, Facultad de Ingeniería, Universidad de Antioquia UdeA; Calle 70 N° 52-21, Medellín, Colombia. Teléfono (574) 2196571.
*Autor de correspondencia: ingry.gomez@udea.edu.co

A review of multivariate statistical methods for analysing water quality

ABSTRACT

In aquatic ecosystems is monitored the water to determine their space-temporary variations, generating large and complex arrays of data that require tools that assist in the interpretation thereof, for managers of water resources can inform society the deterioration of these and take corrective action. This review is a revision of multivariate statistical techniques used to examine the spatial and temporal variability of water quality. We consider few techniques like Factor Analysis, which is used in order to reduce the dimensionality of the data and build underlying latent variables or factors that produce the observed variables, these factors can be used as water quality indexes built from the data collected; we also consider cluster and discriminant analysis than is commonly used to study the spatial variability and studying similarities between periods or sampling stations, these three techniques are commonly used for exploratory purposes; for more complex goals such as modeling and prediction, hierarchical models, Multiple Regression and Structural Equations are presented. For all methods, we present their functionality and usability methods and illustrated using case studies. This review describes how these methods can be used in order to study water quality for monitoring spatial and temporal variability of the measures taken.

Keywords: multivariate statistical methods, structural equations models, hierarchical models, multivariate multiple regression, water quality.

RESUMEN

En los ecosistemas acuáticos se monitorea el agua para determinar sus variaciones espacio-temporales, generando grandes y complejas matrices de datos que requieren herramientas que ayuden en la interpretación de los mismos, para que los administradores de los recursos hídricos puedan informar a la sociedad el deterioro de estos y tomar medidas

Editor: Hernández Fernández, J.

Citation: Gómez, I. & Peñuela, G. (2016). Revisión de los métodos estadísticos multivariados usados en el análisis de calidad de aguas. *Revista Mutis 6*(1), 54-63, doi: http://dx.doi.org/10.21789/22561498.1112

Received: September 7, 2015. **Accepted:** March 7, 2016. **Published on line:** May 31, 2016.

Competing Interests: The authors have no conflict of interest.

correctivas. El presente artículo es una revisión de tema cuyo objetivo es el examen de técnicas estadísticas multivariadas usadas para examinar la variabilidad espacio-temporal de la calidad del agua. En él se presentan diversas técnicas como el análisis factorial, que se usa con el fin de disminuir la dimensionalidad de los datos y construir factores subyacentes o variables latentes que generen las variables observadas, estos factores pueden usarse e interpretarse como índices de calidad del agua construidos a partir de los datos recolectados; también se presenta en análisis de clúster y el análisis discriminante que se usan comúnmente para estudiar la variabilidad espacial, estudiando similaridades entre períodos o estaciones de muestreo, estas tres técnicas se usan comúnmente con fines exploratorios; para objetivos más complejos como el modelamiento y la predicción, se presentan los modelos jerárquicos, de regresión múltiple y de ecuaciones estructurales. Para todos los métodos se presenta su funcionalidad y aplicabilidad y se ilustran usando casos de estudio. Esta revisión describe cómo estos métodos pueden utilizarse con miras a estudiar la calidad del agua con el fin de monitorear espacial y temporalmente la variabilidad de las medidas tomadas.

Palabras clave: métodos estadísticos multivariados, modelos de ecuaciones estructurales, modelos jerárquicos, modelos de regresión múltiple multivariada, calidad de aguas.

INTRODUCCIÓN

El agua es un recurso escaso e indispensable para la supervivencia humana y de la mayoría de las especies en el planeta, es por ello que se hace necesario administrar eficientemente el recurso hídrico y estudiar los impactos que han generado las actividades antropogénicas y los cambios ambientales que ocurren en las cuencas hídricas, lo que se logra estudiando la calidad del agua. La calidad del agua involucra muchos parámetros que pueden variar espacial y temporalmente, de acuerdo a los vertimientos y cambios climáticos. Estos parámetros están relacionados con la presencia de diferentes contaminantes en el agua, disueltos o en suspensión, en concentraciones que varían a lo largo del recurso hídrico por los vertimientos, las transformaciones bióticas o abióticas, sedimentación, etc. Las condiciones ambientales del agua como pH, potencial redox, temperatura, oxígeno disuelto y conductividad eléctrica pueden favorecer la transformación biótica o abiótica de los contaminantes. Igualmente ocurre con las condiciones hidráulicas del recurso hídrico que pueden favorecer la sedimentación de los contaminantes. Un contaminante puede estar evaluado por uno o más parámetros, y un parámetro puede evaluar uno o más tipos de contaminantes; por esto, varios parámetros están relacionados. Por lo tanto, los parámetros de calidad de aguas pueden variar de un sitio a otro y de un día a otro, y por esto, para la interpretación de los datos que se obtengan de los diferentes parámetros en un recurso hídrico, se hace imperante contar con métodos de análisis de datos que permitan evaluar, de manera simultánea, las múltiples relaciones que existen entre las variables (parámetros) y su evolución espacial y temporal, papel que cumplen a cabalidad los métodos estadísticos multivariados.

La aplicación de diferentes métodos estadísticos multivariados como análisis de clúster (Clúster Analysis CA), análisis de componentes principales (Principal Component Analysis PCA), Análisis Factorial (Factor Analysis FA), y análisis discriminante (Discriminant Analysis DA), son de gran ayuda en la interpretación de matrices de datos complejas para un mejor entendimiento de la calidad del agua, que permiten la identificación de posibles factores o fuentes que afectan los sistemas acuáticos y ofrecen una valiosa herramienta para la administración confiable de los recursos hídricos así como soluciones rápidas a los problemas de contaminación (Shrestha & Kazama, 2007). También existen otros métodos estadísticos que estudian las relaciones de causalidad y dependencia, como son el análisis de correlación canónica (Canonic Correlation Analysis CCA), los modelos jerárquicos (Hierarchical Models), modelos de regresión múltiple multivariada (Multiple Multivariate Regession Models) y los modelos de ecuaciones estructurales (Structural Equation Models SEM). El análisis de correlación canónica es ampliamente usado en calidad de aguas con el fin de estudiar las relaciones entre grupos de parámetros, entregando dos vectores de variables, uno que representa las variables endógenas y otro las exógenas, con la particularidad de que la correlación entre estos vectores es máxima, se ha usado en calidad de aguas por ejemplo estudiando las relaciones entre los parámetros físicos (vector

que representa las variables exógenas) y químicos (vector que representa las variables endógenas) (Noori *et al.*, 2010); los modelos jerárquicos, también conocidos como modelos de efectos mixtos (Mixed Effects Models LMM), se usan con el fin de estudiar la correlación espacial y temporal en caso de ser usados en serie de tiempo, para las medidas repetidas, son especialmente útiles cuando se tienen estaciones de muestreo que funcionan de manera independiente; la regresión múltiple multivariada permite estudiar las relaciones de causalidad y dependencia cuando se tiene un conjunto de variables endógenas *versus* otro conjunto de variables exógenas. Los modelos de regresión múltiple han sido usados para encontrar ecuaciones que permiten predecir o controlar variables que afectan la calidad del agua como, por ejemplo, los sólidos disueltos totales (Chenini & Khemiri, 2009); los modelos de ecuaciones estructurales permiten estudiar, de manera simultánea, las relaciones evaluadas en la regresión múltiple y las relaciones entre las variables observadas (endógenas y exógenas) y factores no observados o latentes, constituyéndose en una combinación del análisis factorial y la regresión múltiple, muy útil para estudiar por completo un ecosistema (Grace, *et al.*, 2010), estudiar la contaminación del agua en un embalse (Liu *et al*., 1997), o la calidad del agua en un río (Zou & Yu, 1994), entre otros.

El propósito del presente documento es la revisión de técnicas estadísticas multivariadas usadas para examinar la variabilidad espacio-temporal de la calidad del agua. Se presentan las diferencias entre las técnicas, sus ventajas y limitaciones, así como algunas aplicaciones con el fin de identificar las técnicas más apropiadas para diferentes circunstancias.

En la primera parte se presentan brevemente los diferentes métodos, su definición, objetivos y tipos, la comparación entre ellos y la formulación matemática. En la segunda se encuentran algunas aplicaciones de los métodos y algunos errores encontrados en ellas y la comparación entre los métodos. Finalmente, se tienen las conclusiones y las referencias bibliográficas.

MÉTODOS ESTADÍSTICOS MULTIVARIADOS EN LA CALIDAD DEL AGUA

Los métodos estadísticos multivariados son una herramienta muy útil al momento de evaluar múltiples relaciones de manera simultánea en bases de datos de alta complejidad. Es por ello que son de gran utilidad para el modelamiento en casi todas las áreas, porque permiten un acercamiento a los fenómenos de estudio, tanto en calidad de aguas, como en otras áreas de las ciencias ambientales y demás ciencias del conocimiento. El análisis multivariado consiste en una colección de métodos que pueden ser usados cuando se realizan varias mediciones a diversos individuos u objetos en una o más muestras. Las medidas son conocidas como variables y los individuos u objetos como unidades u observaciones (Rencher, 2003).

Los métodos multivariados priman sobre los univariados porque estos últimos están limitados a examinar uno solo o, a lo sumo, unos pocos procesos al tiempo, estos métodos han predominado en los últimos 50 años y no tienen en cuenta las interacciones en los fenómenos o sistemas bajo estudio (Grace, 2006), por lo tanto, los métodos multivariados trascienden la mirada univariada.

Los métodos multivariados más usados en el análisis de calidad de aguas se dividen en: métodos de reducción de dimensión, métodos de agrupamiento, análisis de clasificación, modelos de regresión múltiple, análisis de correlación canónica, modelos jerárquicos y los modelos de ecuaciones estructurales.

Reducción de dimensión

Las variables que se analizan en calidad de aguas pueden presentar, adicional a las estructuras de dependencia entre grupos de variables, estructuras de correlación dentro de estos grupos, lo que viola el supuesto de independencia que existe en la mayoría de los métodos multivariados. Para darle solución a esta situación y cumpliendo el principio de parsimonia en estadística (dar una explicación con la mayor cantidad de información con el menor número de variables), existen los métodos de reducción de dimensión, principalmente el análisis de componentes principales (PCA) y el análisis factorial (FA).

Análisis de componentes principales (PCA)

El PCA es un procedimiento matemático que transforma un conjunto de variables correlacionadas en un conjunto menor de variables no correlacionadas llamadas componentes principales. Esta técnica tiene dos objetivos: (1) Reducción de dimensión y (2) Facilitar la interpretación de los datos (Johnson & Wichern, 2002). En el análisis de componentes principales se busca maximizar la varianza de una combinación lineal de variables, con la menor pérdida de información posible.

Suponga que **X** es un vector aleatorio de $px1$ con matriz de varianzas covarianzas $\mathbf{\Sigma}_{pxp}$, entonces:

$$Y_i = a_i^T X \quad (1)$$

Donde Y_i es la i-ésima componente principal de **X** y a_i es el i-ésimo vector propio de **Σ**.

Análisis factorial (FA)

El FA tiene por objeto explicar un conjunto de variables observadas por un pequeño número de variables *latentes* o no observadas llamadas *factores* (Peña, 2002). En el análisis factorial se representan las variables $X_1, X_2, \cdots, X_p$ como combinaciones lineales de un pequeño conjunto de variables aleatorias $f_1, f_2, \cdots, f_m$ (con m<<p), llamadas factores, donde los factores son *constructos latentes* que generan las X'_2. El modelo factorial sería:

$$(X-\mu) = LF + \varepsilon \quad (2)$$

Donde **μ** es el vector de medias de **X**, **L** , es la matriz que contiene las cargas o pesos del j-ésimo factor f_j en la i-ésima variable X_1, **F** es la matriz que contiene los factores y **ε** es el vector de errores que da cuenta de la parte de la variable que es única (no común con otras variables).

Las diferencias entre PCA y FA son señaladas por Rencher (2003) y Peña (2002) así:

1. Los componentes principales están definidos como combinaciones lineales de las variables originales. En FA, las variables originales son expresadas como combinaciones lineales de los factores.

2. En PCA se explica una gran parte de la varianza total de las variables. En FA se busca dar cuenta de las covarianzas o correlaciones entre las variables.

3. PCA es una herramienta descriptiva, mientras que FA presupone un modelo estadístico formal de generación de la muestra dada.

Adicionalmente,

4. En PCA los componentes son ortogonales. En FA existen varias metodologías para estimar los factores, si se cumplen los supuestos del método, se puede estimar el modelo ortogonal usando el método de máxima verosimilitud.

5. En FA los ejes se pueden rotar usando rotación varimax para capturar más varianza. En PCA se pueden rotar pero esto no garantiza que se esté capturando más varianza.

6. Los factores son invariantes ante transformaciones lineales de las variables. PCA es muy sensible a estas transformaciones, incluso si se estandarizan las variables, los componentes son diferentes.

Agrupamiento y clasificación

Los métodos de agrupamiento tienen por objeto agrupar unidades experimentales en grupos homogéneos en función de las similaridades entre ellas, estos métodos se conocen como métodos de *aprendizaje no supervisado* porque son las unidades experimentales las que deciden la manera en la que conforman los grupos homogéneos. Los métodos de clasificación se denominan de *aprendizaje supervisado* porque los grupos están conformados *a priori*, y el método confirma qué tan buenos son los agrupamientos, qué variables son las que los determinan y en qué grupo se ubicaría una nueva unidad experimental.

Métodos de agrupamiento: análisis de clúster

El agrupamiento se realiza aplicando el análisis de clúster, cuyo objetivo es encontrar un agrupamiento óptimo en el cual las observaciones u objetos dentro de cada clúster (grupo) son similares, pero los clústeres son diferentes unos de otros (Rencher, 2003). Existen dos enfoques para el análisis de clúster: el jerárquico y el no jerárquico (Johnson, 1998), y básicamente dependen del objetivo del investigador. El más usual es el jerárquico, observándose el agrupamiento en un gráfico llamado dendograma.

Análisis de clasificación: análisis discriminante (DA)

El análisis discriminante (DA) es una técnica multivariada usada para determinar las variables responsables de la separación de las unidades dentro de los grupos (Bierman *et al.*, 2011)which are based on in situ data collection and hence are often spatially or temporally limited. Remote sensing imagery is increasingly used as a rich source of spatial information, providing more detailed coverage then other methods. But the complexity of information in the imagery requires new analysis techniques that allow us to identify the components and possible causes of spatial and temporal variability. This paper presents a review of methods to analyse spatial and temporal variations in remote sensing data of coastal water quality and discusses and compares these methods and the outcomes they achieve. Selected techniques are illustrated by using a sample dataset of MODIS chlorophyll-a imagery. We consider classification methods (cluster analysis, discriminant analysis. El análisis discriminante puede resolver una serie de preguntas, entre las cuales están determinar si hay diferencias estadísticamente significativas entre dos o más grupos conocidos, estableciendo cuáles variables independientes aportan a las diferencias entre los grupos, y encontrar procedimientos para clasificar unidades dentro de ellos (Hair, 2010). El DA se puede considerar como el análisis de regresión en la que la variable endógena "Y" es categórica, que toma valores o categorías para cada grupo, y las variables exógenas son las variables continuas que determinan a qué grupo pertenecen las unidades.

Los métodos de agrupamiento y clasificación son métodos complementarios. Generalmente, se usan de manera simultánea. Inicialmente se usa el CA, jerárquico en la mayoría de los casos debido a la interpretabilidad y visualización del dendograma, para agrupar las unidades y, para estos clústeres establecidos, se aplica el DA con el fin de confirmar los agrupamientos, conocer qué variables discriminan mejor entre grupos y, en caso de incorporar una nueva unidad experimental, una estación de muestreo por ejemplo, saber a cuál de los grupos pertenece.

Otros métodos multivariados

Entre otros métodos multivariados están los de regresión lineal múltiple multivariada y los jerárquicos, que permiten encontrar ecuaciones que describen las relaciones de dependencia en un conjunto de variables. Los primeros son usados cuando se busca un modelo que describa las relaciones de dependencia entre varias variables endógenas explicadas por un conjunto de variables endógenas. El resultado es un modelo del tipo:

$$Y=X\beta+e \qquad (3)$$

Donde **Y** es la matriz que contiene las variables endógenas (explicadas), **X** contiene las variables exógenas (explicatorias), **β** contiene los pesos o contribuciones marginales que tienen las Xs en las Ys, y **e** es el vector de errores.

Los modelos jerárquicos determinan qué variables exógenas predicen o explican mejor la variable endógena y sus interacciones. En estos modelos y otros, basados en el análisis de varianza ANOVA, a las variables exógenas se les llama factores y a sus categorías se les llama niveles. Los modelos jerárquicos usan datos cuya estructura es jerárquica, es decir, las unidades (de observación o experimentales) del primer factor se encuentran anidadas dentro de las unidades del segundo factor, las unidades del segundo factor anidadas en las del tercero y así sucesivamente, por tanto, los parámetros de estos modelos pueden ser visualizados como una estructura lineal jerárquica (Raudenbush & Bryk, 2002).

Los modelos jerárquicos comparados con los modelos de regresión lineal, tienen la ventaja de recoger tanto la variabilidad espacial como la temporal de medidas repetidas para todos los tipos de datos (Pätynen *et al.*, 2013). Mientras que las regresiones lineales tienen la ventaja de que pueden mezclar variables exógenas tanto discretas como continuas, y hasta categóricas, en contraste los modelos jerárquicos no pueden incorporar variables exógenas continuas directamente (Raudenbush & Bryk, 2002).

Modelos de ecuaciones estructurales (SEM)

Los modelos de ecuaciones estructurales, SEM, relacionan estados de entrada, procesos y salidas a través de variables exógenas y endógenas, en los que las variables endógenas pueden convertirse en variables de exógenas en otros momentos. Esta técnica estima relaciones de dependencia múltiples y cruzadas, que incorpora conceptos no observados, llamados constructos. La mejor forma de determinar el modelo de

ecuaciones estructurales es a través de la gráfica denominada diagrama de secuencias, debido a que en esta las relaciones de interdependencia se representan a través de flechas directas que señalan el impacto o causalidad de la variable exógena sobre la variable endógena y, las flechas curvadas señalan la correlación entre las variables. Las ecuaciones del modelo de ecuaciones estructurales son:

$$\eta=\beta\eta+\Gamma\xi+\varsigma \quad (4)$$

Donde:

η es el vector de variables aleatorias latentes endógenas, **ξ** el vector de variables aleatorias latentes exógenas, **β** la matriz de coeficientes entre latentes dependientes, **Γ** la matriz de coeficientes entre variables latentes dependientes e independientes y **ς** vector de perturbaciones.

$$x=\Lambda_x \xi+\delta_x \quad (5)$$

$$y=\Lambda_y\eta+\varepsilon_x \quad (6)$$

Donde:

x es el vector de p variables observadas, $\mathbf{\Lambda_x}$ la matriz de coeficientes que muestran las relaciones entre las variables latentes y observadas exógenas, **ξ** la latente exógena y **δ** el vector de errores asociados a las variables exógenas.

y el vector de q variables observadas, $\mathbf{\Lambda_y}$ la matriz de coeficientes que muestran las relaciones entre las variables latentes y observadas endógenas, **η** la latente endógena y el vector de errores asociados a variables endógenas.

La ecuación (4) se conoce como el modelo estructural y las ecuaciones (5) y (6) como los modelos de medida.

Validación de supuestos

En estadística, los supuestos son condiciones que se deben cumplir para que los modelos puedan usarse de manera confiable y no llegar a conclusiones erradas en la interpretación de los mismos. En PCA no existen los supuestos porque es un método matemático pero no estadístico, en cambio en el FA sí existen supuestos y se deben hacer validaciones (Johnson & Wichern, 2002); en el análisis de clúster, independiente del método elegido, se deben cumplir los supuestos de normalidad e independencia de las variables (Johnson & Wichern, 2002; Rencher, 2003); los supuestos de la regresión múltiple multivariada son los mismos de la regresión múltiple, (Gujarati, 1988); tanto en los modelos jerárquicos (Montgomery, 2008) como en SEM (Hair, 2010) existen supuestos.

APLICACIONES EN CALIDAD DE AGUAS

Los métodos estadísticos multivariados han sido ampliamente usados desde la década de 1980 debido al desarrollo de paquetes estadísticos que facilitan el modelamiento matemático. En calidad de aguas se encuentran aplicaciones con todos los métodos mencionados previamente.

La construcción de índices de calidad de aguas es una de las principales aplicaciones de los métodos de reducción de dimensión, por ejemplo Coletti *et al.* (2010), construyeron un índice de calidad del agua aplicando el análisis factorial para determinar la influencia de las actividades agrícolas en la calidad del recurso hídrico del río Das Pedras ubicado en las regiones de Mogi Guaçu y Estiva Gerbi en Brasil; para ello se analizaron los siguientes parámetros de calidad de aguas: conductividad eléctrica, pH, nitrógeno amoniacal, amonio, nitratos, fósforo total, sólidos suspendidos, turbiedad y oxígeno disuelto, durante trece meses, demostrándose, a partir de dicho índice, que la calidad del agua en el río Das Pedras se ha deteriorado progresivamente debido a las actividades agrícolas.

Los modelos SEM se han usado para proponer valores estándar de los parámetros que inciden en la eutrofización con el fin de apoyar los procesos de legislación y establecer límites en una ecoregión de China (Ji *et al*., 2013)there has been no nutrient standard established for LE control in many developing countries such as China. This study proposes a structural equation model to assist in the establishment of a lake nutrient standard for drinking water sources in Yunnan-Guizhou Plateau Ecoregion (Yungui Ecoregion; la metodología partió con una consulta a expertos que arrojó una serie de modelos estructurales evaluando datos históricos por más de diez (10) años, determinándose que el fósforo total y la clorofila a fueron las variables determinantes en la eutrofización, inclusive fijaron valores límites para estas. De igual manera, Grace *et al.* (2010), but also because of its promise as a means of representing

theoretical concepts using latent variables. In this paper, we discuss characteristics of ecological theory and some of the challenges for proper specification of theoretical ideas in structural equation models (SE models desarrollaron dos modelos SEM; en el primero pretendieron estudiar la relación entre la recuperación de la vegetación después de una conflagración con la edad de las plantas y la severidad del incendio, en el segundo tenían como objetivo predecir el conteo de especies en un río con base en cuatro variables latentes (estrés abiótico, disturbancia, biomasa y diversidad de plantas); adicionalmente, los investigadores presentaron propuestas de desarrollos teóricos en los SEM. Grace (2008) hace una brevísima descripción de los SEM, análisis de dos casos de estudio, descripción de la notación LISREL (Linear Structural Relations) y un breve recuento histórico del desarrollo de los SEM.

Aplicación de los métodos estadísticos

De otro lado, se ha realizado la combinación de varias técnicas estadísticas con el fin de enriquecer la interpretación de los fenómenos estudiados y complementar los aportes de las técnicas individuales.

Métodos de reducción de dimensión, agrupamiento y clasificación

El análisis de componentes principales se usó, acompañado del análisis de correlación canónica, CCA (Noori *et al*., 2010), para estudiar el río Karoon en Irán. Se monitorearon 12 parámetros entre los que estaban la turbiedad, los sólidos suspendidos totales, la demanda química de oxígeno, los sulfatos y los nitratos. El objetivo fue identificar las estaciones de muestreo más significativas para el monitoreo de la calidad del agua y los parámetros de calidad más importantes, mediante el PCA, y relacionar los parámetros físicos *versus* los químicos en dicho río, mediante el CCA. Aplicando el análisis de componentes principales se descartaron 4 de las 17 estaciones de muestreo y se determinó una correlación de 0.993 entre los parámetros físicos y químicos.

Bierman *et al*. (2011), realizaron una revisión bibliográfica del uso combinado de PCA, FA, CA y DA para el monitoreo de la calidad del agua en zonas costeras. Los autores recomiendan el uso de los métodos de agrupamiento y clasificación (CA y DA) tanto en el análisis exploratorio, confirmatorio como predictivo, porque resumen e identifican patrones en bases de datos altamente complejas, mientras que recomiendan el uso de PCA y FA para identificar relaciones entre las variables, identificar variables representativas de un conjunto grande de ellas y crear un conjunto menor de variables que reemplace las originales en análisis posteriores. También recomiendan usar mapas autoorganizados (Self-organising maps-SOM) que son una forma de redes neuronales (no son métodos estadísticos), para extraer patrones en grandes conjuntos de datos, y los semivariogramas para obtener una medida de la variabilidad entre mediciones, conforme su separación espacial se incrementa. Shrestha & Kazama (2007), aplicaron estos métodos en la cuenca del río Fuji en Japón para extraer información acerca de las similaridades o disimilaridades entre sitios de muestreo usando CA, identificación de las variables responsables de las variaciones espaciales y temporales en la calidad del agua del río usando DA, determinación de los factores subyacentes que explican la estructura de la base de datos usando PCA y FA, y la influencia de posibles fuentes antropogénicas en los parámetros de calidad del agua. Se midieron 12 parámetros de calidad del agua en 13 sitios de muestreo a lo largo del río durante las 4 estaciones del año. Los resultados a los que llegaron son: los principales parámetros responsables de la calidad del agua son los relacionados con las descargas al río, la temperatura ambiente y la contaminación orgánica; la cuenca del río se clasifica, según su grado de contaminación, en tres áreas, alta, media y baja, cuyo grado de contaminación es debido principalmente a las fuentes antropogénicas como las aguas residuales domésticas, los fertilizantes y las industrias. También Varol *et al.* (2012) aplicaron estas metodologías estadísticas en la cuenca del río Tigris en Turquía.

Modelos SEM y otros métodos estadísticos

Los modelos SEM se usan combinados con técnicas como PCA, FA, regresión múltiple, los modelos jerárquicos y el análisis de series de tiempo. Zou & Yu (1994), desarrollaron un modelo SEM para modelar la calidad del agua en el río Arkansas en Estados Unidos, monitoreando 14 parámetros mensualmente durante 14 años en una estación de muestreo; el modelo desarrollado se combinó con el análisis factorial para determinar el número de constructos a usar (se eligieron cinco). Obtuvieron cinco modelos que se compararon tanto desde el punto de vista teórico como técnico con el fin de elegir aquel que cumpliera los supues-

tos y que mejor describiera las interacciones entre los diferentes parámetros. Zou & Yu (1994), muestran la validez de los supuestos.

De igual manera, Wu *et al.* (2014which was first conducted to determine four types of factors, respectively, those for organic pollution, eutrophication, seasonal influence, and sediment pollution. The analysis results effectively help to determine water quality in the watershed of the reservoir. The authors reutilize analysis of moment structures (AMOS) combinaron los SEM y el análisis factorial, para construir un conjunto estándar de métodos que pudieran usar las autoridades que manejan los embalses, para mejorar la calidad de aguas, tanto en las cuencas como en los embalses. Wu *et al.* (2014) usaron 6 estaciones de monitoreo a lo largo de la reserva Fetsui en Taiwán, en las que se midieron 9 parámetros de calidad de agua: pH, temperatura, oxígeno disuelto, demanda bioquímica de oxígeno, sólidos suspendidos, surfactantes aniónicos, nitrógeno amoniacal, fósforo total, y clorofila_a durante dos años y medio. Comprobaron que las 9 variables son generadas por 4 constructos: contaminación orgánica, eutrofización, estacionalidad y contaminación por sedimentos. Se evaluaron tres modelos estructurales a partir de estas variables, pero uno de ellos (modelo 3) fue el más adecuado para el propósito inicial. Esta aplicación presenta el inconveniente de que no valida los supuestos.

La calidad de las aguas subterráneas también se ha estudiado usando los SEM combinados con los modelos de regresión, el análisis de componentes principales y el análisis de clúster. Un ejemplo de ello es el trabajo de Chenini & Khemiri (2009) quienes estudiaron un área de 1250 km^2 localizados en la región de Atlas en Túnez, en la cual se monitoreó un sistema de tres acuíferos tomando 28 muestras de agua, midiendo 10 parámetros entre octubre y noviembre de 2005. Iniciaron con un PCA para determinar las relaciones entre las propiedades del agua analizadas e identificar los factores que afectan la concentración de cada uno, encontrándose tres componentes principales que acumulan el 70 % de la varianza total; seguidamente realizaron CA en las variables para conocer las semejanzas entre ellas; luego se encuentra un modelo usando la regresión lineal para predecir los sólidos disueltos totales a partir de los valores de magnesio, calcio, sodio, cloro, HCO_3^- y el SO_4^{2-}; y, finalmente, construyeron un modelo estructural, SEM, para predecir de manera simultánea los sólidos disueltos totales y el cloro, con base en el magnesio, calcio, sodio, HCO_3^- y el SO_4^{2-}, con el fin de proveer una explicación adecuada de las interacciones simultáneas de las variables en el modelo conceptual.

Otra aplicación en aguas subterráneas, la realizaron Liu *et al*. (1997), quienes usaron los modelos SEM y las series de tiempo para investigar la influencia del clima, la hidrología y la dosificación de nitrógeno en la producción agrícola, sobre el área de Big Spring en Iowa, Estados Unidos. Los investigadores eligieron como variables endógenas la concentración de nitrógeno y las descargas; como variables exógenas la precipitación, la temperatura del aire, la evapotranspiración potencial y el balance de nitrógeno, que se midieron mensualmente entre 1982 y 1991. En el modelamiento se incluyó un rezago de un período de tiempo para las variables endógenas, incorporando las series de tiempo en el modelo SEM. Se ajustó un modelo para cada una de las cuatro estaciones del año: verano, invierno, primavera y otoño, comprobando que la influencia del clima, la producción agrícola y la tendencia creciente en el tiempo de la concentración de nitrógeno en el suelo, son los factores que más afectan la dinámica de la contaminación por nitrógeno en el agua subterránea del área bajo estudio.

Pätynen *et al*. (2013), muestran la aplicación de los modelos SEM y los modelos jerárquicos en el estudio de la ecología acuática, indicando que estos modelos son una importante alternativa para evaluar la interacción entre diferentes variables, conocer la incidencia de factores subyacentes que pueden explicar diversas situaciones en el modelamiento ecológico, modelar espacialmente y desarrollar la creatividad (principalmente en el uso de los SEM). Señalan desventajas de estos modelos, como la gran cantidad de datos que requieren, que no siempre se pueden obtener desde el punto de vista técnico, y la alta capacidad de las herramientas computacionales para el modelamiento.

Comparación de métodos

Los métodos expuestos, más que ser rivales y que el investigador tenga que decidir cuál usar, son complementarios. Los métodos de reducción de dimensión FA y PCA se pueden usar si se incumple el supuesto de independencia de las variables, sin embargo, se debe elegir uno de ellos, se recomienda el uso de FA cuando las variables sean normales

(se pueden usar tranformaciones de potencia para normalizar) y se apliquen otros métodos estadísticos posteriormente.

En análisis de clúster se puede realizar de dos formas, jerárquica y no jerárquica. El CA jerárquico se recomienda para tamaños de muestra pequeños dado que el objetivo es conocer las similaridades entre las variables o individuos, el clúster no jerárquico es el adecuado para grandes tamaños de muestra, en este caso observar la separación de los grupos es lo primordial. Generalmente las aplicaciones de CA van seguidas de DA, con el fin de conocer las variables responsables de la separación entre los grupos (o clústeres); y también, aunque menos común, clasificar unidades (estaciones de muestreo por ejemplo) que no habían sido tenidas en cuenta en la observación inicial, en los grupos ya existentes. El CA y DA tienen los supuestos de normalidad e independencia de las variables, lo que provoca que la confiabilidad en los resultados se vea limitada, incluso restringida, cuando no se cumplen.

Los métodos de FA, PCA, CA y DA, son para fines exploratorios. Mientras que el modelamiento usando modelos SEM, jerárquicos, de regresión y demás, son para análisis más profundos que involucran descripción, control y predicción de las relaciones de asociación y dependencia entre las variables.

Los modelos de regresión son una gran familia que agrupa todo el modelamiento estadístico basado en el ANOVA. Sin lugar a dudas son los modelos estadísticos más usados en todas las áreas del conocimiento, comprenden la regresión simple, múltiple, múltiple multivariada, los modelos lineales generalizados (MLG), las series de tiempo, y demás; su gran ventaja es que entregan una o varias ecuaciones que sirven, entre otros, para obtener pronósticos confiables. Tienen los supuestos de normalidad, independencia de las variables exógenas, varianza constante (homocedasticidad) y no autocorrelación de los residuales; el no cumplimiento de estos supuestos presenta diagnósticos y pronósticos poco confiables.

Los SEM son una combinación del análisis factorial confirmatorio y los modelos de regresión, por ello tienen la gran ventaja de que permiten modelar simultáneamente múltiples relaciones de causalidad y dependencia; sin embargo, requieren un conocimiento *a priori* del fenómeno estudiado y las posibles relaciones entre las variables, debido a que los SEM responden a fines confirmatorios más que exploratorios. Además, tienen fuertes supuestos de normalidad, cointegración y demás, propios de los modelos estadísticos complejos.

El investigador puede usar las técnicas exploratorias y de modelación estadística de manera simultánea en el análisis de un mismo fenómeno, esto le permitirá tener una visión más completa del comportamiento univariado, multivariado y de las relaciones entre las variables.

CONCLUSIONES

Los métodos estadísticos multivariados son herramientas muy valiosas en los estudios de la calidad del agua. Permiten reducir la dimensionalidad de los datos, determinar factores subyacentes que generen las variables involucradas en los estudios, conocer las variaciones espaciales y temporales de las dinámicas presentes en los cuerpos de agua, obtener modelos que permiten evaluar las relaciones entre las variables de manera simultánea y, principalmente, apoyar la toma de decisiones mediante el diagnóstico y predicción de los fenómenos estudiados. Estos métodos pueden complementarse con otros análisis como los de redes neuronales y los geoestadísticos para tener una visión más amplia de los fenómenos bajo estudio.

En la actualidad existe la tendencia en la aplicación de los modelos de ecuaciones estructurales en la calidad de aguas, esto se debe a su gran potencia al evaluar las relaciones de entrada, los procesos y las salidas a través de variables exógenas y endógenas, en los que las variables endógenas pueden convertirse en variables exógenas en otros momentos, e incorpora conceptos no observados, llamados constructos, lo que permite un mayor y mejor acercamiento al comportamiento de las dinámicas en los cuerpos de agua.

En varias de las aplicaciones se encontró que no se validan los supuestos, lo cual puede ocasionar una falta grave a la confiabilidad de los modelos. Se recomienda que siempre que se use un método estadístico se validen los supuestos del modelo, en caso de violarse alguno o varios de ellos, se deben hacer los procedimientos a que haya lugar para corregirlos.

Es de resaltar que estos métodos requieren de un alto poder computacional y de altos volúmenes de datos, lo que en muchas ocasiones no es viable económica y técnicamente.

REFERENCIAS

Abbasi, T., & Abbasi, S. A. (2012). *Water Quality Indices*. Elsevier Science.

Bierman, P., Lewis, M., Ostendorf, B., & Tanner, J. (2011). A review of methods for analysing spatial and temporal patterns in coastal water quality. *Ecological Indicators*, *11*(1), 103-114.

Chenini, I., & Khemiri, S. (2009). Evaluation of ground water quality using multiple linear regression and structural equation modeling. *International Journal of Environmental Science & Technology*, *6*(3), 509-519.

Coletti, C., Testezlaf, R., Ribeiro, T. A. P., Souza, R. T. G. de, & Pereira, D. de A. (2010). Water quality index using multivariate factorial analysis. *Revista Brasileira de Engenharia Agrícola e Ambiental*, *14*(2), 517-522.

Grace, J. B. (2006). *Structural Equation Modeling and Natural Systems*. Cambridge University Press, Ed.

Grace, J. B. (2008). Structural Equation Modeling for Observational Studies. *The Journal of Wildlife Management*, *72*(1), 14-22.

Grace, J. B., Anderson, T. M., Olff, H., & Scheiner, S. M. (2010). On the specification of structural equation models for ecological systems. *Ecological Monographs*, *80*(1), 67-87.

Gujarati, D. N. (1988). *Basic Econometrics*. Mc-Graw-Hill.

Hair, J. F. (2010). *Multivariate Data Analysis* (7th ed.). Prentice Hall.

Ji, D., Xi, B., Su, J., Huo, S., He, L., Liu, H., & Yang, Q. (2013). A model to determine the lake nutrient standards for drinking water sources in Yunnan-Guizhou Plateau Ecoregion, China. *Journal of Environmental Sciences*, *25*(9), 1773-1783.

Johnson, D. E. (1998). *Applied Multivariate Methods for Data Analysts*. Duxbury Press.

Johnson, R. A., & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis* (5th ed.). Prentice Hall.

Liu, Z.-J., Hallberg, G. R., & Malanson, G. P. (1997). Structural Equation Modeling of Dynamics of Nitrate Contamination in Ground Water1. *JAWRA Journal of the American Water Resources Association*, *33*(6), 1219-1235.

Montgomery, D. C. (2008). *Design and Analysis of Experiments*. John Wiley & Sons.

Noori, R., Sabahi, M. S., Karbassi, A. R., Baghvand, A., & Taati Zadeh, H. (2010). Multivariate statistical analysis of surface water quality based on correlations and variations in the data set. *Desalination*, *260*(1-3), 129-136.

Pätynen, A., Kotamäki, N., & Malve, O. (2013). Alternative approaches to modelling lake ecosystems. *Freshwater Reviews*, *6*(2), 63-74.

Peña, D. (2002). *Análisis de datos multivariantes* (1st ed.). Mac-Graw Hill.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE Publications.

Rencher, A. C. (2003). *Methods of Multivariate Analysis*. (2003 John Wiley & Sons, Ed.) (second edi). Wiley-Interscience.

Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling & Software*, *22*(4), 464-475.

Varol, M., Gökot, B., Bekleyen, A., & Şen, B. (2012). Spatial and temporal variations in surface water quality of the dam reservoirs in the Tigris River basin, Turkey. *CATENA*, *92*(0), 11-21.

Wu, E., Tsai, C., Cheng, J., Kuo, S., & Lu, W. (2014). The Application of Water Quality Monitoring Data in a Reservoir Watershed Using AMOS Confirmatory Factor Analyses. *Environmental Modeling & Assessment*, *19*(4), 325-333.

Zou, S., & Yu, Y.-S. (1994). A general structural equation model for river water quality data. *Journal of Hydrology*, *162*(1-2), 197-209.