EL LORO HUASTECO
Órgano de Divulgación Científica y Tecnológica del
Instituto Tecnológico Superior de Pánuco

# Análisis estadísticos multivariados en estudios ambientales: ¿La panacea o una caja de pandora?

Carlos Alberto Contreras Verteramo[1*], Reina Verónica Román Salinas[1], Marco Antonio Díaz Martínez[1]

[1]Instituto Tecnológico Superior de Pánuco

*contrerc@itspanuco.edu.mx

Los métodos estadísticos multivariados son una serie de técnicas descriptivas e inferenciales desarrolladas para analizar fenómenos en que intervienen conjuntos de variables que pueden ser predictivas o de respuesta. Estos métodos son apropiados cuando es necesario analizar más de una característica para describir una unidad de muestreo y también cuando las relaciones entre esas características requieren ser analizadas de forma simultánea. Utilizada correctamente, la estadística multivariada puede ser una herramienta eficiente, que permite al investigador transcribir, interpretar y trasmitir mensajes poco evidentes o subyacentes, contenidos en los datos (McGarigal *et al*. 2000).

Las técnicas estadísticas multivariados pueden ser clasificados como descriptivas e inferenciales, pero en ningún momento estas funciones son independientes ya que, al determinar la significancia estadística (función inferencial), automáticamente se conoce la combinación de las variables originales (función descriptiva) que aporta la evidencia de mayor peso para contrastar la hipótesis nula (McGarigal *et al*. 2000; Greenacre y Primicerio 2013).

Los análisis multivariados comprenden una amplia variedad de técnicas extensamente empleadas en estudios ecológicos y ambientales; entre los métodos descriptivos más utilizados se encuentran los de ordenación como el análisis de componentes principales (ACP), análisis de factores (AF), análisis de correspondencia (AC), escalamiento multidimensional (EMM), los de agrupamiento como el análisis de clusters, k-medias y análisis de funciones discriminantes (AFD), así como métodos inferenciales entre los que se pueden mencionar el

análisis de correlación canónica (ACC) y el de marginalidad media (OMI) (McGarigal et al. 2000; Smith 2012).

Mediante el uso correcto del ACP, los autores pueden identificar la o las variables que aportan la mayor variación a los datos (Sánchez-Rojas *et al*. 2009; Ortiz-Martínez *et al*. 2005), describir una matriz de datos multivariados por medio de la reducción de sus dimensiones y visualizarlos gráficamente, encontrar las combinaciones lineales no correlacionadas de las variables originales con la mayor variación (Mesa-Zavala *et al*. 2012) y obtener nuevas variables combinadas para posteriores estudios del sistema.

Es común que algunos autores utilicen erróneamente este método para relacionar o hacer comparaciones (Cabrera y Wallace 2007), ya que únicamente está diseñado para identificar tendencias. Esta técnica también puede utilizarse (deliberadamente o no) para confundir al lector al representar gráficamente componentes principales que explican una proporción mínima de la variación o con figuras en las que no se definen claramente las características de los componentes principales (Castaño 2001; Mesa-Zavala *et al*. 2012).

De acuerdo con Pérez y Medrano (2010), el AF permite simplificar los datos eliminando la redundancia o duplicación de un grupo de variables correlacionadas, representar las variables altamente correlacionadas mediante un menor número de factores "latentes" o "subyacentes" independientes entre sí facilitando la identificación del origen de las diferencias (Riiters *et al*. 1995; Liu *et al*. 2013; Rais *et al*. 2015). Algunos autores omiten reportar algunos resultados del análisis o bien no utilizan los resultados al integrar sus conclusiones (Fulton *et al*. 1996; Bokor *et al*. 2013).

El AC ofrece la posibilidad de considerar múltiples variables categóricas que permiten revelar relaciones imposibles de ser detectadas mediante comparaciones pareadas, muestra la forma en que se relacionan las variables, permite crear diagramas de dispersión biespaciales a la medida, con los que se puede llegar a conclusiones basadas en la interpretación gráfica, es posible asociar directamente frecuencia de especies con los ambientes de los que proceden y además permite prueba de hipótesis, por lo que se considera también una prueba inferencial (Hill y Lewicki 2007; Hair *et al*. 2014).

Comúnmente autores que utilizan el análisis de correspondencia se enfocan en la interpretación de la gráfica de dispersión que arroja la prueba, dejando de lado el reporte de

resultados como Chi cuadrada y los grados de libertad (Jude y Pappas 1992; Knight 1995; De Miguel *et al*. 1997; Williams *et al*. 1997; Díaz-Varela *et al*. 2011) privando al lector de información importante para la evaluación de los resultados presentados en el documento.

El análisis de Escalamiento Multidimensional (EMM) Métrico o No Métrico es un procedimiento multivariado de interdependencia cuyo fin es la identificación de las distancias o disimilitud en una matriz de datos, así como su representación gráfica en un espacio de pocas dimensiones. Es un método de ordenación bastante robusto para reducir las dimensiones de los datos sin hacer transformaciones previas. Se basa en la comparación de datos, de forma qué si A y B son los más parecidos, esta técnica los colocará en el gráfico de manera que la distancia entre ellos sea menor que entre cualquier otro par de datos.

El uso de este análisis se justifica para la reducción de las dimensiones de la información debido a que soporta la ausencia de normalidad multivariada y es muy robusto ante alta frecuencia de ceros en los datos como se puede apreciar en los trabajos de Backhaus *et al*. (1987), Laurance y Laurance (1999) o Muhly *et al*. (2011). En algunos artículos los autores no reportan los valores de stress, el número de iteraciones realizadas o la proporción de la varianza representada por cada eje (Laurance y Laurance 1999; Mata *et al*. 2008; Schultz *et al*. 2011), limitando la comprensión del análisis y comprometiendo la calidad de su trabajo.

Los análisis de agrupamiento (clusters y k-medias) son un conjunto de procedimientos analíticos cuyo propósito principal es descubrir agrupaciones significativas basándose en un gran número de variables interdependientes, siempre y cuando se especifique el algoritmo de amalgamación y las medidas de distancia utilizadas, como lo presentan Amstrup *et al*. (2004) y Zharikov *et al*. (2005) en sus respectivos trabajos. En contraste algunos autores cometen errores como la utilización de medidas de distancia incorrectas para el tipo de datos colectados, lo que por consecuencia invalida o desacredita sus resultados (McFarlane y Boxall 1996; Muhly *et al*. 2011; Bager y Fontoura 2013).

El AFD comprende dos procedimientos relacionados cuyo principal objetivo es describir las diferencias entre dos o más grupos definidos, así como predecir la posibilidad de que un dato de origen desconocido pertenezca a un grupo particular, con base en una variable agrupadora categórica, cuando el análisis es realizado y presentado de manera correcta se reportan los valores de Wilk´s Lambda (WL), F, g.l. y p además de las gráficas de dispersión, como es el caso de los trabajos de Kitchener *et al*. (2006), Reed *et al*. (2006) y Smallwood *et al*. (1993).

En algunos casos los autores omiten presentar los resultados de los análisis de manera correcta, lo que origina confusión en el lector y pérdida de confianza en la veracidad del reporte presentado (Morrison 1984; Caughley 1987).

El análisis de correlación canónica puede ser considerado una extensión de la regresión múltiple y de la correlación, se aplica a situaciones donde es apropiada la técnica de la regresión, pero para más de una variable dependiente. Aunque otra aplicación del análisis de correlación canónica es como un método para determinar la asociación entre dos grupos de variables. Cuando esta técnica es utilizada de manera correcta, es posible describir relaciones complejas entre dos grupos de variables (Williams *et al*. 2002; Rodewald y Bakermans 2006). En contraste, cuando no se reportan debidamente los resultados del análisis se complica la explicación y disminuye la confianza del lector hacia el trabajo (Boyce 1978; Brandt *et al*. 2000; Suárez *et al*. 2004).

El análisis de marginalidad media representa el promedio canónico de las condiciones ambientales del área de estudio (centro de gravedad) y la desviación que a partir de éste presenta el promedio de las condiciones ambientales utilizadas por una especie (centroide). Por lo tanto, se asume que las especies con altos valores de marginalidad están influidas por un subgrupo de las variables ambientales registradas, mientras que los valores bajos indican que la especie no responden específicamente a algunas variables ambientales, éstas especies tienden a estar presentes a lo largo de toda el área de estudio; la tolerancia representa la amplitud de las condiciones en la que se presenta la especie; por último, la tolerancia residual se refiere a la variación que no es explicada por la marginalidad ni por la tolerancia. Al utilizar esta técnica estadística de manera correcta, es posible identificar la sensibilidad de las especies a las variables ambientales a lo largo de gradientes (Randa y Yunger 2006; Sorensen *et al*. 2015). En cambio, si los resultados del análisis no son reportados correctamente u omitidos, es posible confundir al lector generando una incertidumbre sobre el conocimiento del autor acerca de la prueba específicamente (Broennimann *et al*. 2006; Hurlbert y White 2007; Dufour *et al*. 2015).

La complejidad de los estudios ecológicos ha ido en aumento en los últimos años y las técnicas estadísticas multivariadas se han convertido en herramientas de gran utilidad para resolver situaciones excepcionalmente difíciles de analizar, que involucran un gran número de variables; pero el sólo hecho de añadir variables a los análisis no resuelve la necesidad de

conocimientos científicos que permitan comprender el problema, es imperativo que quienes aplican métodos estadísticos multivariados comprendan que no son un buen sustituto de la ignorancia acerca de los sistemas biológicos o ecológicos. Es importante no perder de vista los requerimientos matemáticos y supuestos de las distintas técnicas multivariadas, pero tampoco se debe olvidar al organismo, ya que su biología es la pieza fundamental del fenómeno estudiado (Johnson 1981; McGarigal *et al*. 2000).

Cuando es necesario identificar la forma en que una especie o población responde a distintos factores ambientales surge la necesidad de un enfoque multivariado, siendo necesario considerar que no todos los parámetros ambientales potencialmente importantes pueden ser medidos; muchas de las variables registradas pueden estar correlacionadas, ser relativamente irrelevantes o irrelevantes para el problema que se busca resolver; o bien, que las variables analizadas y potencialmente relevantes darán como resultado un conjunto de datos multidimensionales de complicada interpretación (Green 1971).

Existen diversas explicaciones para el mal uso de los métodos estadísticos multivariados que se pueden observar en los trabajos científicos publicados alrededor del planeta, entre ellos se puede mencionar que los autores desconocen las diversas técnicas y su aplicación, que sólo reportan los resultados de técnicas multivariadas cuando coinciden con su intuición, conocimiento previo o resultados de análisis univariados. En diversos trabajos los métodos multivariados no tienen realmente utilidad, sólo son presentados para producir un efecto de credibilidad sobre conclusiones alcanzadas mediante técnicas más simples (Johnson 1981).

Los análisis multivariados son herramientas muy útiles para resolver interrogantes ecológicas de gran complejidad y cuando se utilizan correctamente, pueden ser la panacea; pero la realidad es que en gran número de publicaciones los supuestos no se cumplen, los autores cometen errores u omisiones en su aplicación o simplemente los resultados no se reportan correctamente, generando problemas en lugar de respuestas correctas y conocimiento científico, convirtiendo una gran herramienta para el análisis de datos en una caja de Pandora.

## Literatura citada

Amstrup, S., T. McDonald, and G. Durner. 2004. Using satellite radiotelemetry data to delineate and manage wildlifepopulations. Wildlife Society Bulletin 32(3):661–679.

Backhaus, W., R. Menzel, and S. Kreibl. 1987. Multidimensional Scaling of Color Similarity in Bees. Biological Cybernetics 56(5):293-304.

Bokor, J., A. Bokor, J. Nagy, P. Horn and I. Nagy. 2013. Summary of Comments on Analysis of Hungarian red deer trophies by means of Principal component analysis in two different counties. Journal of Central European Agriculture 14(1):452-466.

Boyce, M. S. 1978. Climatic variability and body size variation in the muskrats (Ondatra zibethicus) of North America. Oecologia 36:1-19.

Brandt, L., K. Portier and W. Kitchens. 2000. Patterns of change in tree islands in Arthur R. Marshall Loxahatchee national wildlife refuge from 1950 to 1991. Wetlands 20(1):1–14.

Broennimann, O., W. Thuiller, G. Hughes, G. Midgley, R. Alkemades and A. Guisan. 2006. Do geographic distribution, niche property and life form explain plants' vulnerability to global change? Global Change Biology 12:1079–1093.

Cabrera, W. H. y R. Wallace. 2007. Densidad y distribución espacial de palmeras arborescentes en un bosque preandino-amazónico de Bolivia. Ecología en Bolivia 42(2):121-135.

Castaño, G. J. 2001. Evaluación de la avifauna asociada a humedales costeros de la Guajira con fines de conservación. Crónica Forestal y del Medio Ambiente 16:5-33.

Caughley, G., J. Short, G. Grigg and H. Nix. 1987. Kangaroos and climate: an analisys of distribution. Journal of Animal Ecology 56:751-761.

Dufour, C., C. Meynard, J. Watson, C. Rioux, S. Benhamou, J. Perez, J. du Plessis, N. Avenant, N. Pillay and G. Ganem. 2015. Space use variation in co-occurring sister species: response to environmental variation or competition? PLoSONE 10(2): e0117750.

Fulton, D., M. J. Manfredo and J. Lipscomb. 1996. Wildlife value orientations: A conceptual and measurement approach. Human Dimensions of Wildlife 1(2):24-47.

Greenacre, M and R. Primicerio. 2013. Multivariate analysis of ecological data. Fundación BBVA, Bilbao, Spain. 331 p.

Hair, J., W. Black, B. Babin and R. Anderson. 2014. Multivariate data analysis. 7a Ed. Pearson Education Ltd. London, UK. 761 p.

Hill, T. and Lewicki, P. 2007. Statistics: Methods and Applications. Dell, Tulsa, OK, U.S.A.

Hurlbert, A. and E. White. 2007. Ecological correlates of geographic range occupancy in North American birds. Global Ecology and Biogeography 16:764-773.

Johnson, D. H. 1981. The use and misuse of statistics in wildlife habitat studies. In: Capen, E. (Ed). The Use of Multivariate Statistics in Studies of Wildlife Habitat. University of Vermont - U.S. Fish and Wildlife Service. 249 p.

Jude, D. J. and J. Pappas. 1992. Fish utilization of great lakes coastal wetlands. J. Great Lakes Res. 18(4):651-672.

Kitchener, A. C., M. Beaumont and D. Richardson. 2006. Geographical variation in the clouded leopard, Neofelis nebulosa, reveals two species. Current Biology 16:2377–2383.

Knight, M. H. 1995. Drought-related mortality of wildlife in the southern Kalahari and the role of man. Afr. J. Ecol. 33:377-394.

Laurance, S. G., and W. F. Laurance.1999. Tropical wildlife corridors: use of linear rainforest remnants by arboreal mammals. Biological Conservation 91:231-239.

Liu Y., M. Zhang and J. Ma. 2013. Correlations of environmental factors with the roe deer (Capreolus pygargus) population genetic variability. International Journal of Digital Content Technology and its Applications 7(6):400-407.

Mata, C., I. Hervás, J. Herranz, F. Suárez, J.E. Malo. 2008. Are motorway wildlife passages worth building? Vertebrate use of road-crossing structures on a Spanish motorway. Journal of Environmental Management 88:407–415.

McFarlane, B. and P. Boxall. 1996. Participation in Wildlife Conservation by Birdwatchers. Human Dimensions of Wildlife 1(3):1-14.

McGarigal, K., S. Cushman and S. Stafford. 2000. Multivariate statistics for wildlife and ecology research. Springer Science + Business Media, LLC. New York, U.S.A. 283 p.

Mesa-Zavala, E., S. Álvarez-Cárdenas, P. Galina-Tessaro, E. Troyo-Diéguez e I. Guerrero-Cárdenas. 2012. Vertebrados terrestres registrados mediante foto-trampeo en arroyos estacionales y cañadas con agua superficial en un hábitat semiárido de Baja California Sur, México. Revista Mexicana de Biodiversidad 83:235-245.

Morrison, M. 1984. Influence of simple size on discriminant function analisys of habitat use by birds. J. Field Ornithol. 55(3)330-335.

Muhly T., C. Semeniuk, A. Massolo, L. Hickman and M. Musiani. 2011. Human activity helps prey win the predator-prey space race. PLoS ONE 6(3):e17050. doi:10.1371/journal. pone.0017050

Nath, S., S. Marcus y B. Druss. 2006. Retractions in the research literature: misconduct or mistakes? Med. J. Aust. 185:152–154.

Ortiz-Martínez, T., S. Gallina, M. Briones-Salas y G. González. 2005. Densidad poblacional y caracterización del hábitat del venado cola blanca (Odocoileus virginianus oaxacensis, Goldman y Kellog, 1940) en un bosque templado de la sierra norte de Oaxaca, México. Acta Zoológica Mexicana (n.s.) 21(3):65-78.

Pérez, E. R. y L. Medrano. 2010. Análisis Factorial Exploratorio: Bases Conceptuales y Metodológicas. Revista Argentina de Ciencias del Comportamiento 2(1):58-66.

Rais, M., A. Akram, S. M. Ali, M. A. Asadi, M. Jahangir, M. J. Jilani and M. Anwar. 2015. Qualitative analysis of factors influencing the diversity and spatial distribution of herpetofauna in Chakwal Tehsil (Chakwal District), Punjab, Pakistan. Herpetological Conservation and Biology 10(3):801–810.

L. Randa and J. Yunger. 2006. Carnivore occurrence along an urban–rural gradient: a landscape-level analysis. Journal of Mammalogy 87(6):1154–1164.

Reed, J. E., R. Baker, W. Ballard and B. Kelly. 2004. Differentiating Mexican gray wolf and coyote scats using DNA analysis. Wildlife Society Bulletin 32(3):685–692.

Rodewald, A. and M. Bakermans. 2006. What is the appropriate paradigm for riparian forest conservation? Biological Conservation 128:193–200.

Schultz, R., S. Andrews, L. O'Reilly, V. Bouchard and S. Frey. 2011. Plant Community Composition More Predictive than Diversity of Carbon Cycling in Freshwater Wetlands. Wetlands 31:965–977.

Smith, L. I. 2002. A tutorial on principal components analysis. University of Otago. New Zealand. 26 p.

Sánchez-Rojas, G., C. Aguilar-Miguel y E. Hernández-Cid. 2009. Estudio poblacional y uso de hábitat por el venado cola blanca (*Odocoileus virginianus*) en un bosque templado de la Sierra de Pachuca, Hidalgo, México. Tropical Conservation Science 2(2):204-214.

Smallwood, K. S. and E. L. Fitzhugh. 1993. A rigorous technique for identifying individual mountain lions *Felis concolor* by their tracks. Biological Conservation 65:51-59.

Sorensen, A., F. van Beest and R. Brook. 2015. Quantifying overlap in crop selection patterns among three sympatric ungulates in an agricultural landscape. Basic and Applied Ecology 16(7):601-609.

Suárez, Y., M. Júnior and A. Catella. 2004. Factors regulating diversity and abundance of fish communities in Pantanal lagoons, Brazil. Fisheries Management and Ecology 11:45–50.

Williams, R., A. Trites and D. Bain. 2002. Behavioural responses of killer whales (*Orcinus orca*) to whale-watching boats: opportunistic observations and experimental approaches. J. Zool. Lond. 256:255-270.

Zharikov, Y., G. Skilleter, N. Loneragan, T. Taranto and B. Cameron. 2005. Mapping and characterising subtropical estuarine landscapes using aerial photography and GIS for potential application in wildlife conservation and management. Biological Conservation 125:87–100.