Revista Bases de la Ciencia - ISSN 2588-0764

Publicación Cuatrimestral. Vol. 4, No 2, Mayo/Agosto, 2019, Ecuador (p. 51-72)

Ciencias Matemáticas

Artículo de Investigación

ASPECTOS TEÓRICOS DEL MANOVA-BIPLOT Y SU APLICACIÓN A LOS USOS DEL SUELO EN LA ESPAC-2016

Dr. Isidro Amaro[1], Dr. Ernesto Ponsot[*], Ph.D. Zenaida Castillo[2], M.Sc. Wilfre Machado[3]

Universidad de Investigación de Tecnología Experimental Yachay. Escuela de Ciencias Matemáticas y Computacionales. Ecuador.

*Autor para correspondencia: eponsot@yachaytec.edu.ec

Recibido: 26-02-2019 / Aceptado: 20-05-2019 / Publicación: 30-05-2019

Editor Académico: Dr. Luis Sánchez

RESUMEN

Se presentan algunos aspectos teóricos sobre la relación entre las técnicas clásicas de MANOVA y Biplot, que son la base de la gráfica MANOVA-Biplot. Se aplica esta técnica a un conjunto de datos referentes al uso del suelo en las tres regiones definidas por el Ecuador. Se indagan sus diferencias con el propósito de lograr una caracterización sintética que sirva para compararlas estadísticamente y que sea de utilidad al proceso de toma de decisiones. La región oriental está caracterizada por el uso de su suelo preponderantemente en montes, bosques naturales y artificiales, la región costera por cultivos permanentes, transitorios y barbechos, y la región de la sierra por los restantes pastos, descansos, páramos y otros. Se presenta una ilustración de los resultados que muestra el gran valor de síntesis que tiene la técnica estadística propuesta, incluido el código R que la produce.

Palabras clave: Regiones del Ecuador, usos del suelo, MANOVA, Biplot.

THEORETICAL ASPECTS OF THE MANOVA-BIPLOT AND THEIR APPLICATION TO LAND USES IN THE ESPAC-2016

ABSTRACT

Some theoretical aspects about the relationship between the classic techniques of MANOVA and Biplot, which are the basis of the MANOVA - Biplot graph are presented. This technique is applied to a set of data concerning to land uses in the the three major regions defined by Ecuador. Their differences are investigated in order to achieve a synthetic characterization that can be used to compare them statistically and to improve appropriately the decision process. The eastern region is characterized by the use of its soil predominantly in mountains, natural and artificial forests, the coastal region by permanent crops, transients and fallows, and The Sierra region by the permanent and transient crops and fallows. An illustration about the results showing the great synthesis value that the proposed technique has, including the R code that produces it is also presented.

Key words: Regions of Ecuador, land uses, MANOVA, Biplot.

ASPECTOS TEÓRICOS DA MANOVA-BIPLOT E SUA APLICAÇÃO AOS USOS DA TERRA NA ESPAC-2016

RESUMO

Alguns aspectos teóricos são apresentados sobre a relação entre as técnicas clássicas de MANOVA e Biplot, que são a base do gráfico MANOVA - Biplot. Aplicando essa técnica a um conjunto de dados referentes aos usos da terra nas três principais regiões definidas pelo Equador, suas diferenças são investigadas para obter uma caracterização sintética que possa ser usada para compará-las estatisticamente e melhorar adequadamente o processo de decisão. . A região leste é caracterizada pelo uso de seu solo predominantemente em montanhas, florestas naturais e artificiais, enquanto a região costeira por culturas permanentes, transiente e favelas, e a região da serra pelos pastos remanescentes, quebras, terras estéreis e outras. É apresentada uma ilustração dos resultados que mostram o grande valor de síntese que a técnica estatística proposta possui, incluindo o código R que a produz.

Palavras-chave: Regiões do Equador, usos do solo, MANOVA, Biplot.

Citación sugerida: Amaro, I., Pensot, E., Castillo, Z., Machado, W. (2019). Aspectos teóricos del Manova-Biplot y su aplicación a los usos del suelo en la ESPAC-2016. Revista Bases de la Ciencia, 4(2), 51-72. DOI: https://doi.org/10.33936/rev_bas_de_la_ciencia.v4i2.1668 Recuperado de: https://revistas.utm.edu.ec/index.php/Basedelaciencia/article/view/1668

Orcid IDs:

Isidro Amaro: https://orcid.org/0000-0003-2402-910X
Ernesto Ponsot: https://orcid.org/0000-0001-5221-1799
Zenaida Castillo: https://orcid.org/0000-0002-4424-8652
Wilfre Machado: https://orcid.org/0000-0002-3797-2159
Luis Sánchez: https://orcid.org/0000-0002-1850-0631

1. INTRODUCCIÓN

El Manova-Biplot es una técnica estadística multivariante utilizada en situaciones experimentales donde se dispone de varias variables respuesta y se quiere buscar las diferencias entre varios grupos. Fue denominada Manova-Biplot por Gabriel (1972, 1995), aunque también se le conoce como Biplots Canónicos según Vicente-Villardón (1992) y Gower Y Hand (1996). El Manova-Biplot para diseños de dos vías fue introducido por Amaro et al. (2004).

Esta útil herramienta, además de permitir la interpretación de las diferencias-semejanzas entre grupos, que es el objetivo principal del Análisis de Varianza Multivariante (MANOVA, por sus siglas en inglés), también ofrece la posibilidad de visualizar las relaciones entre las variables, y las relaciones entre grupos y variables, al mismo tiempo que proporciona medidas de calidad de representación tanto para variables como para medias de grupos, lo que facilita significativamente la interpretación de los resultados. Una descripción más detallada de la teoría de los métodos Biplot puede encontrarse en Nieto et al. (2014). Algunos usos recientes del MANOVA-Biplot se discuten en Iñigo et al. (2004), Varas et al. (2005), Iñigo et al. (2014), García-Talegón et al. (2016) e Iñigo et al. (2017).

En este trabajo se presentan algunos aspectos teóricos de la relación entre MANOVA y Biplot, que es la base para la construcción del MANOVA-Biplot como técnica conjunta, así como para probar sus propiedades estadísticas. Se propone usar estos principios en la construcción de regiones elípticas de confianza, en lugar de las clásicas regiones circulares. La teoría propuesta se soporta con programas elaborados en R (R Core Team 2017). Se presenta como ilustración de la técnica, un estudio de las características de las regiones que componen el Ecuador en términos del uso de su suelo. Estas características tienen implicaciones en el proceso de decisiones sobre qué usos incentivar. También influyen en las políticas de apoyo a la economía productiva que potencian la calidad de vida de los habitantes.

Específicamente, a partir de los datos contenidos en la Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC 2016), elaborada por el Instituto Nacional de Estadística y Censos (INEC) del Ecuador (INEC 2017), se emplea la técnica MANOVA-Biplot para construir una gráfica sintética de las variables (usos de suelo) y grupos (regiones). Un análisis de la gráfica generada permite apreciar las calidades de representación y la variabilidad explicada. Se describen las características resaltantes de las regiones examinadas, mostrando la gran valía de la técnica estadística para sintetizar la información contenida en 3567 registros de la región Sierra, 2069 de la región Costa y 931 de la región Oriental.

2. Metodología

En esta sección se describen los principales elementos teóricos que han sido utilizados para el estudio propuesto. Se exponen las características del MANOVA y su relación con el MANOVA-Biplot, en la construcción automática de elipses de confianza, para el análisis de un conjunto de datos multivariantes.

2.1. Análisis de Varianza Multivariante (MANOVA)

Cuando se dispone de un conjunto de p variables de respuestas independientes observadas para n individuos, como herramienta de análisis simultaneo del conjunto, el MANOVA propone el modelo lineal:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E} \qquad (1)$$

En (1), $\boldsymbol{Y}_{(nxp)} = [y_{ij}]$ es la matriz de datos, cuyas columnas se forman con las variables de respuesta observadas, $\boldsymbol{X}_{(nxm)} = [x_{ij}]$ es la matriz de diseño, decidida por el investigador, $\boldsymbol{B}_{(mxp)} = [\beta_{ij}]$ la matriz de parámetros a estimar y $\boldsymbol{E}_{(nxp)} = [\epsilon_{ij}]$, la matriz de desviaciones aleatorias.

Supóngase además que las filas de $\boldsymbol{E}$ se distribuyen como normales independientes, esto es $\boldsymbol{\epsilon}_{ij} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ con $\boldsymbol{\epsilon}_i = [\epsilon_{i1} \ \ \epsilon_{i1} \ \ldots \ \epsilon_{ip}]'$ para $i = 1, 2, \ldots, n$ y $\boldsymbol{\Sigma}_{(pxp)}$ su matriz de varianzas y covarianzas (Cuadras 2014, 265).

Se puede demostrar que si $\boldsymbol{X}$ es una matriz de rango completo por columnas, el estimador de mínimos cuadrados ordinarios de $\boldsymbol{B}$ es $\widehat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ y $\widehat{\boldsymbol{\Sigma}} = [\boldsymbol{R}_0/(n-m)]$ con $\boldsymbol{R}_0 = \boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y}$

Una vez ajustado, este modelo es útil para probar hipótesis lineales del tipo:

$$H_0: \boldsymbol{CBM} = 0 \qquad (2)$$

con $\boldsymbol{C}$ y $\boldsymbol{M}$ matrices de dimensiones apropiadas y rango completo, compuestas por constantes conocidas.

ISNN 2588-0764

Se busca la caracterización de grupos a partir de variables que se encuentran identificadas como pertenecientes a dichos grupos *a priori*. En particular, si se ordenan los elementos de datos en función de los grupos de pertenencia, el procedimiento supone la matriz de diseño en la forma siguiente:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Esto es, con unos en las filas que se correspondan con el *j*-ésimo grupo, representados estos últimos por las columnas de $\boldsymbol{X}$ para $j = 1, 2, \cdots, m$. Nótese también que con la configuración mostrada, la matriz $\boldsymbol{X}$ resulta de rango completo por columnas.

2.2. MANOVA-Biplot

El MANOVA-Biplot combina criterios de métodos multivariantes clásicos para construir una mejor técnica que permite analizar matrices de datos, proporcionando ejes con máximo poder discriminante y representaciones simultáneas de poblaciones y variables en el mismo sistema de referencia.

Amaro et al. (2004) desarrollaron la teoría necesaria para la construcción del MANOVA- Biplot en el contexto MANOVA de dos vías. A partir de $\widehat{\boldsymbol{B}}$, utilizando la descomposición en valores singulares de la matriz $\boldsymbol{C}\widehat{\boldsymbol{B}}$, el método procura obtener la representación de variables y grupos en una misma figura, determinando las calidades de representación de ambos conjuntos en el menor número de ejes posibles, de preferencia, dos ejes.

Suponiendo el modelo presentado en (1) y la hipótesis propuesta en (2), un Biplot para la matriz $\widehat{\boldsymbol{D}} = \boldsymbol{C}\widehat{\boldsymbol{B}}\boldsymbol{M} = \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{M}$ se puede construir a partir de la descomposición en valores singulares

generalizada[1] como:

$$\boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{D}}\boldsymbol{R}_0^{-\frac{1}{2}} = \boldsymbol{U}\boldsymbol{\Delta}_\lambda\boldsymbol{V}' \tag{3}$$

donde $\boldsymbol{W} = \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'$, $\boldsymbol{R}_0$ es la matriz de suma de cuadrados y productos dentro de grupos (no singular) y $\boldsymbol{\Delta}_\lambda$ es la matriz diagonal de los valores propios de la matriz correspondiente.

La relación entre MANOVA y Biplot se produce debido a que los λ's que aparecen en la descomposición (3), son los mismos valores singulares de la matriz $\boldsymbol{H}\boldsymbol{R}_0^{-1}$, utilizados en MANOVA, con

$$\begin{aligned}\boldsymbol{H} &= \widehat{\boldsymbol{D}}'\boldsymbol{W}^{-1}\widehat{\boldsymbol{D}}\\ &= \boldsymbol{M}'\boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\boldsymbol{W}^{-1}\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{M}\\ &= \boldsymbol{M}'\boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1}\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{M}\end{aligned}$$

Esto se demuestra en el siguiente teorema:

Teorema 1. *Los valores propios incluidos en* $\boldsymbol{\Delta}_\lambda$ *de* (3) *son también valores propios de la matriz* $\boldsymbol{H}\boldsymbol{R}_0^{-1}$.

Demostración: De la descomposición (3), considerando λ el valor propio correspondiente al vector propio $\boldsymbol{v}$ se tiene que:

$$\left\{\left(\boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{D}}\boldsymbol{R}_0^{-\frac{1}{2}}\right)'\left(\boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{D}}\boldsymbol{R}_0^{-\frac{1}{2}}\right) - \lambda\boldsymbol{I}\right\}\boldsymbol{v} = 0$$

Dado que $\boldsymbol{W}$ y $\boldsymbol{R}_0$ son matrices simétricas:

$$\left\{\boldsymbol{R}_0^{-\frac{1}{2}}\widehat{\boldsymbol{D}}'\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{D}}\boldsymbol{R}_0^{-\frac{1}{2}} - \lambda\boldsymbol{I}\right\}\boldsymbol{v} = 0$$

$$\left\{\boldsymbol{R}_0^{-\frac{1}{2}}\widehat{\boldsymbol{D}}'\boldsymbol{W}^{-1}\widehat{\boldsymbol{D}}\boldsymbol{R}_0^{-\frac{1}{2}} - \lambda\boldsymbol{I}\right\}\boldsymbol{v} = 0$$

[1] Se utiliza la definición de raíz cuadrada de una matriz como $\boldsymbol{A}^{1/2} = \boldsymbol{U}\sqrt{\boldsymbol{D}}\boldsymbol{U}^{-1}$ donde $\boldsymbol{D}$ es la matriz diagonal con los valores singulares de $\boldsymbol{A}$, ordenados de mayor a menor y $\boldsymbol{U}$ contiene los vectores singulares.

$$\left\{\boldsymbol{R}_0^{-\frac{1}{2}}\widehat{\boldsymbol{D}}'\boldsymbol{W}^{-1}\widehat{\boldsymbol{D}}\boldsymbol{R}_0^{-\frac{1}{2}} - \lambda\boldsymbol{R}_0^{-\frac{1}{2}}\boldsymbol{R}_0^{\frac{1}{2}}\right\}\boldsymbol{v} = 0$$

$$\boldsymbol{R}_0^{-\frac{1}{2}}\left\{\widehat{\boldsymbol{D}}'\boldsymbol{W}^{-1}\widehat{\boldsymbol{D}}\boldsymbol{R}_0^{-\frac{1}{2}} - \lambda\boldsymbol{R}_0^{\frac{1}{2}}\right\}\boldsymbol{v} = 0$$

$$\boldsymbol{R}_0^{-\frac{1}{2}}\left\{\widehat{\boldsymbol{D}}'\boldsymbol{W}^{-1}\widehat{\boldsymbol{D}}\boldsymbol{R}_0^{-\frac{1}{2}} - \lambda\boldsymbol{R}_0\boldsymbol{R}_0^{-\frac{1}{2}}\right\}\boldsymbol{v} = 0$$

$$\boldsymbol{R}_0^{-\frac{1}{2}}\left\{\widehat{\boldsymbol{D}}'\boldsymbol{W}^{-1}\widehat{\boldsymbol{D}} - \lambda\boldsymbol{R}_0\right\}\boldsymbol{R}_0^{-\frac{1}{2}}\boldsymbol{v} = 0$$

$$\Leftrightarrow \left\{\widehat{\boldsymbol{D}}'\boldsymbol{W}^{-1}\widehat{\boldsymbol{D}} - \lambda\boldsymbol{R}_0\right\}\boldsymbol{R}_0^{-\frac{1}{2}}\boldsymbol{v} = 0$$

Donde $\widehat{\boldsymbol{D}} = \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{M}$, luego:

$$[\boldsymbol{H} - \lambda\boldsymbol{R}_0]\boldsymbol{R}_0^{-\frac{1}{2}}\boldsymbol{v} = 0$$

$$[\boldsymbol{H} - \lambda\boldsymbol{R}_0]\boldsymbol{R}_0^{-1}\boldsymbol{R}_0^{\frac{1}{2}}\boldsymbol{v} = 0$$

$$\left[\boldsymbol{H}\boldsymbol{R}_0^{-1} - \lambda\boldsymbol{I}\right]\boldsymbol{R}_0^{\frac{1}{2}}\boldsymbol{v} = 0$$

En consecuencia, λ es un valor propio de $\boldsymbol{H}\boldsymbol{R}_0^{-1}$.

En la literatura clásica de MANOVA, ver por ejemplo Mardia et al. (1979) y Morrison (1978), $\boldsymbol{H}$ es la matriz de sumas de cuadrados y productos *entre* grupos. Adicionalmente, los λ son también utilizados en las cuatro pruebas de hipótesis de la técnica, a saber: Lambda de Wilks, Prueba de Lawley-Hotelling, Prueba de Pillai y la prueba de Roy o de la mayor raíz.

2.3. Construyendo elipses de confianza

El procedimiento consiste en ubicar los centroides de grupos en la matriz $\boldsymbol{P} = (\boldsymbol{X}'\boldsymbol{X})^{-1/2}\boldsymbol{U}\boldsymbol{D}$. $\boldsymbol{P}$ es una matriz cuadrada que contiene tantas filas como grupos haya definido el investigador (m). Su j-ésima columna representa la coordenada en el eje j para $j = 1, 2, \cdots, m$.

Por otra parte, el procedimiento entrega las coordenadas de las variables en la matriz $\boldsymbol{Q} = \boldsymbol{R}_0^{1/2}\boldsymbol{V}$, usando las definiciones previas. Análogamente a lo que sucede con la matriz $\boldsymbol{P}$, $\boldsymbol{Q}$ es

una matriz cuadrada que contiene tantas filas como variables (p). Su j-ésima columna representa la coordenada en el eje j para $j = 1, 2, \cdots, p$.

En el contexto bidimensional (considerando solo los dos primeros ejes), para lograr una región de confianza alrededor de los centroides de grupos, en contraposición a la clásica región circular, Amaro et al. (2008) proponen regiones elípticas del $(1 - \alpha)\%$ de confianza, construidas usando la matriz de dispersión de dichos centroides $\boldsymbol{S}_{(2x2)}$, como sigue:

$$(\boldsymbol{x}_i - \boldsymbol{p}_i)'\boldsymbol{S}^{-1}(\boldsymbol{x}_i - \boldsymbol{p}_i) = \chi^2_{(2,\alpha)}, \qquad i = 1,2,\ldots,k$$

Donde $\boldsymbol{p}_i$ es la i-ésima fila de la matriz $\boldsymbol{P}$ y $\chi^2_{(2,\alpha)}$ es el valor crítico de nivel α de una distribución χ-cuadrada con dos grados de libertad. La forma cuadrática, dispuesta al lado izquierdo de la ecuación, representa una elipse rotada y trasladada, cuyo despliegue gráfico aprovecha los valores y vectores propios de la matriz $\boldsymbol{S}^{-1}/\chi^2_{(2,\alpha)}$ para la determinación de sus semiejes principales y el ángulo de rotación.

3. Aplicación a los usos del suelo en la ESPAC 2016

En esta sección se exponen los resultados obtenidos mediante la aplicación de las técnicas de MANOVA-Biplot y elipses de confianza, en lugar de las clásicas regiones de confianza circulares, sobre el conjunto de los datos recabados en la Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC 2016), elaborada por el Instituto Nacional de Estadística y Censos (INEC) del Ecuador (INEC, 2016).

3.1. Descripción de los datos

Desde el punto de vista computacional, la metodología se basa en los archivos de datos de la ESPAC 2016, elaborada por el INEC en dicho año. Utilizando el capítulo 3 de la encuesta, dedicado al uso de la tierra, a partir de los archivos CSV disponibles, se migró a una base de datos PostgreSQL sobre la cual se aplicaron los factores de expansión, se agregaron variables, se depuraron valores faltantes y relativizaron las cantidades en hectáreas de las variables. El procesamiento estadístico de los datos se realizó en R (R Core Team 2017) y el código programado se presenta como Anexo. La definición de las variables se muestra en la **tabla 1.**

Tabla 1. Variables del estudio.

Variable	Significado	Agregación
cultiv	Suelo destinado a cultivos	Cultivos permanentes, transitorios y barbechos
descan	Suelo en descanso	No
pasto	Suelo destinado a pastos	Pastos naturales y cultivados
param	Suelo de páramos	No
montbos	Suelo destinado a montes y bosques	Montes, bosques naturales y bosques artificiales
otros	Suelo destinado a otros usos	No

Fuente: Elaboración propia a partir de la ESPAC 2016.

La **figura 1** muestra por regiones, las proporciones de registros contenidos en la muestra, así como las proporciones del área total cubierta, en cada caso. La región Sierra es aquella con mayor número de puntos muestrales con 3567 (54,32 %), le sigue la región Costa con 2069 (31,51 %) y finalmente la Región Oriental con 931 (14,18 %). Por otra parte, en cuanto a la superficie total cubierta por la encuesta, la mayor proporción corresponde a la región Costa (39,05 %) por ejemplo. Esto sugiere un menor parcelamiento (y por lo tanto, parcelas de mayor tamaño) en dicha región, cuando se la compara con las restantes.

La **tabla 2** contiene las cifras absolutas y relativas totales de usos de suelo por regiones, incluidas en la ESPAC 2016. Nótese en la tabla 2 que el uso del suelo predominante en la región oriental es montes y bosques con un 78,27 %, mientras que en las regiones Sierra y Costa alcanza apenas al 38,72 % y 28,31 % respectivamente. La región que destina mayor superficie al cultivo es la Costa, mientras que aquella que destina mayor superficie a los pastos es la Sierra.

Las áreas destinadas a descanso son relativamente pequeñas en las tres regiones, por lo cual es de suponer que el Ecuador está haciendo un uso extensivo e intensivo de sus tierras cultivables.

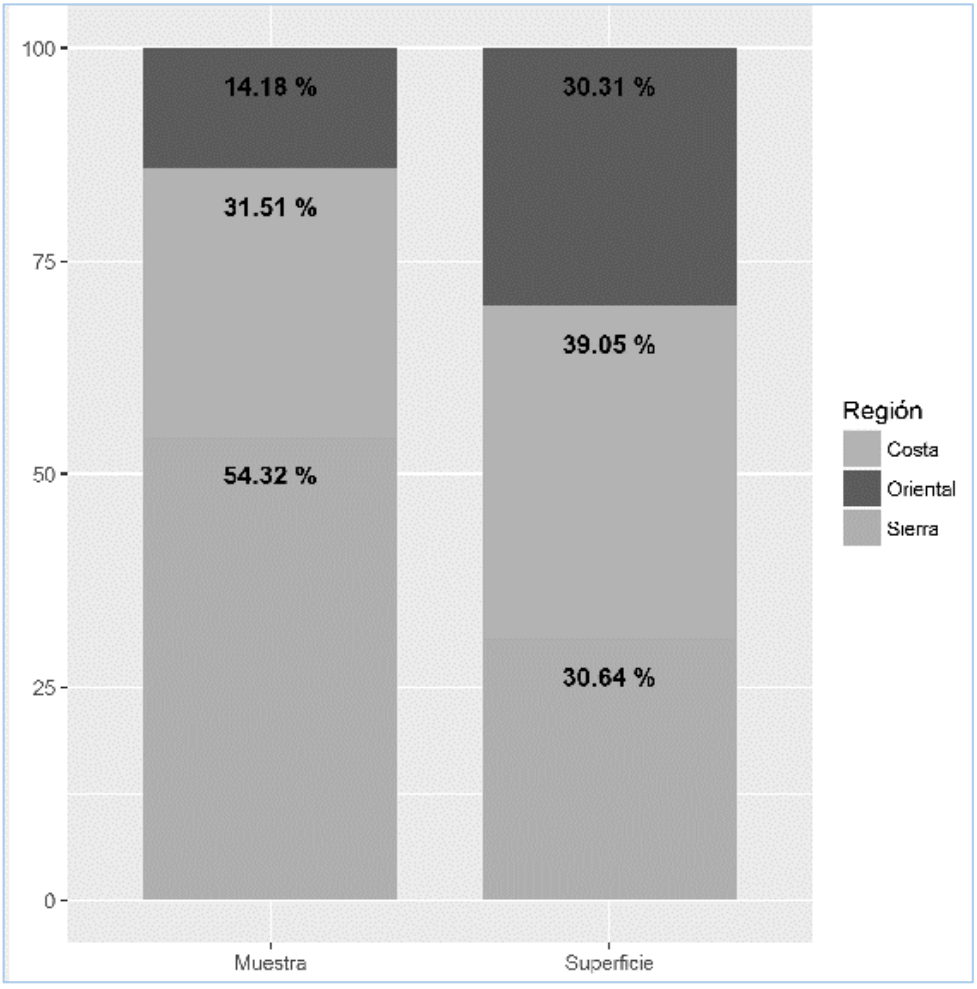

Figura 1. Proporción de elementos en la muestra y cobertura por regiones.

Fuente: Elaboración propia a partir de la ESPAC2016.

Tabla 2. Cifras absolutas y relativas totales de usos de suelo por regiones.

Uso del Suelo	Sierra		Costa		Oriental	
	N°	%	N°	%	N°	%
Cultiv	495.938	13,09	1.632.432	33,80	150.032	4,00
Descan	54.824	1,45	65.384	1,35	5.732	0,15
Pasto	1.235.610	32,61	1.411.680	29,23	451.575	12,05
Param	361.994	9,55	4.996	0,10	10.802	0,29
montbos	1.467.340	38,72	1.367.320	28,31	2.933.670	78,27
Otros	173.801	4,59	348.065	7,21	196.384	5,24
Total	3.789.507	100,00	4.829.877	100.00	3.748.195	100.00

Fuente: Elaboración propia a partir de la ESPAC 2016.

ISNN 2588-0764

La **figura 2** muestra gráficos de caja para cada variable (en términos relativos dentro de cada región), categorizadas por regiones. De la figura se aprecia claramente que las variables descansos, páramos y otros usos, resultan representadas muy marginalmente en la muestra, mientras que las variables cultivos, pastos, montes y bosques resultan ampliamente representadas.

Los cultivos son predominantes en la región Costa, los pastos en la región Sierra y los montes y bosques, en la región Oriental. No obstante, los pastos muestran una variabilidad similar en las tres regiones, mientras que las otras dos variables importantes varían de forma sustantivamente diferente por regiones.

4. DISCUSIÓN

Una vez obtenidos y depurados los datos de la ESPAC 2016 que se refieren a los usos del suelo en el Ecuador, se aplicaron las técnicas MANOVA-Biplot con regiones elípticas de confianza, con los resultados que se discuten a continuación.

En la **tabla 3** se presentan los valores propios, inercias e inercias acumuladas para el Biplot. Allí puede verse que la inercia acumulada casi alcanza el 100% en los dos primeros ejes.

Por otra parte, en la **figura 3** se muestra el MANOVA-Biplot para los grupos, representados por las regiones y las variables seleccionadas, consideradas para el análisis como proporciones de uso del suelo, en cada registro de la encuesta (ver las **tablas 1 y 2**), en lugar de las cifras absolutas (reportadas en hectáreas). Para diferenciar cada variable en términos absolutos de aquella considerada en términos relativos (o proporciones), a los nombres de dichas variables les antecede la letra “p".

En la **figura 3** se puede observar que las tres regiones (Sierra, Costa y Oriente) están claramente diferenciadas en cuanto a las variables de uso del suelo, ya que sus elipses de confianza no se interceptan. La variable que más contribuye a la diferencia de los tres grupos es el porcentaje de cultivos (*pcultiv*), porque al proyectar las elipses de confianza de los grupos sobre el vector que representa a esa variable, las proyecciones son disjuntas.

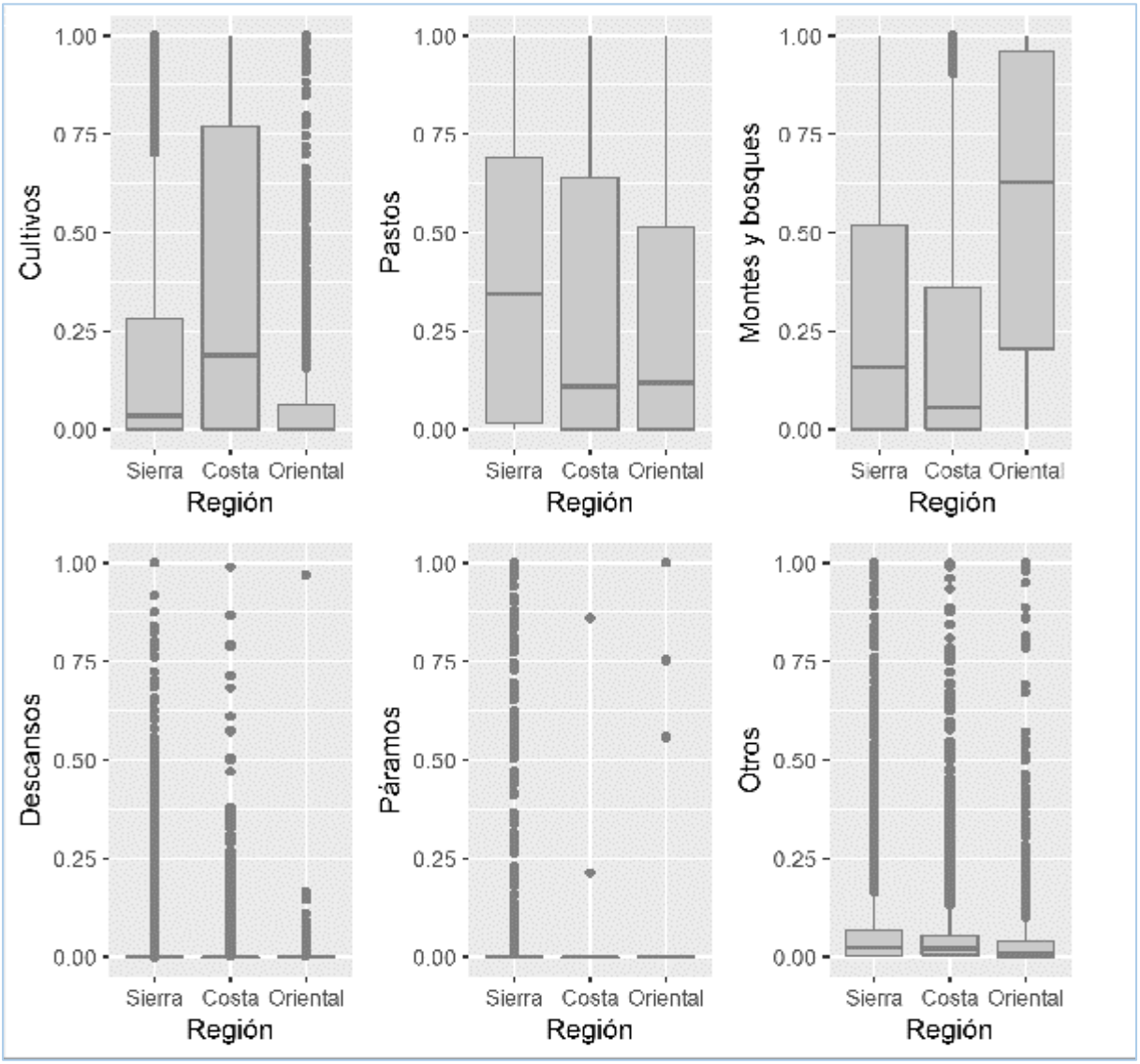

Figura 2. Boxplot de regiones y variables.

Fuente: Elaboración propia a partir de la ESPAC2016.

Tabla 3. Valores propios, inercias e inercias acumuladas.

EJE	Valores propios	Porcentaje de inercia	Inercia acumulada
1	3,680 * 10-1	75,128	75,128
2	2,118 * 10-1	24,870	99,998
3	2,500 * 10-6	0,030	100,000

Fuente: Elaboración propia a partir de la ESPAC 2016.

Por otra parte, la variable que más influye en la separación de los grupos Sierra y Oriente es el porcentaje de montes, bosques naturales y artificiales (pmonbos) pues el vector que la representa es casi paralelo a la recta que pasa por los centroides de dichos grupos.

ISNN 2588-0764

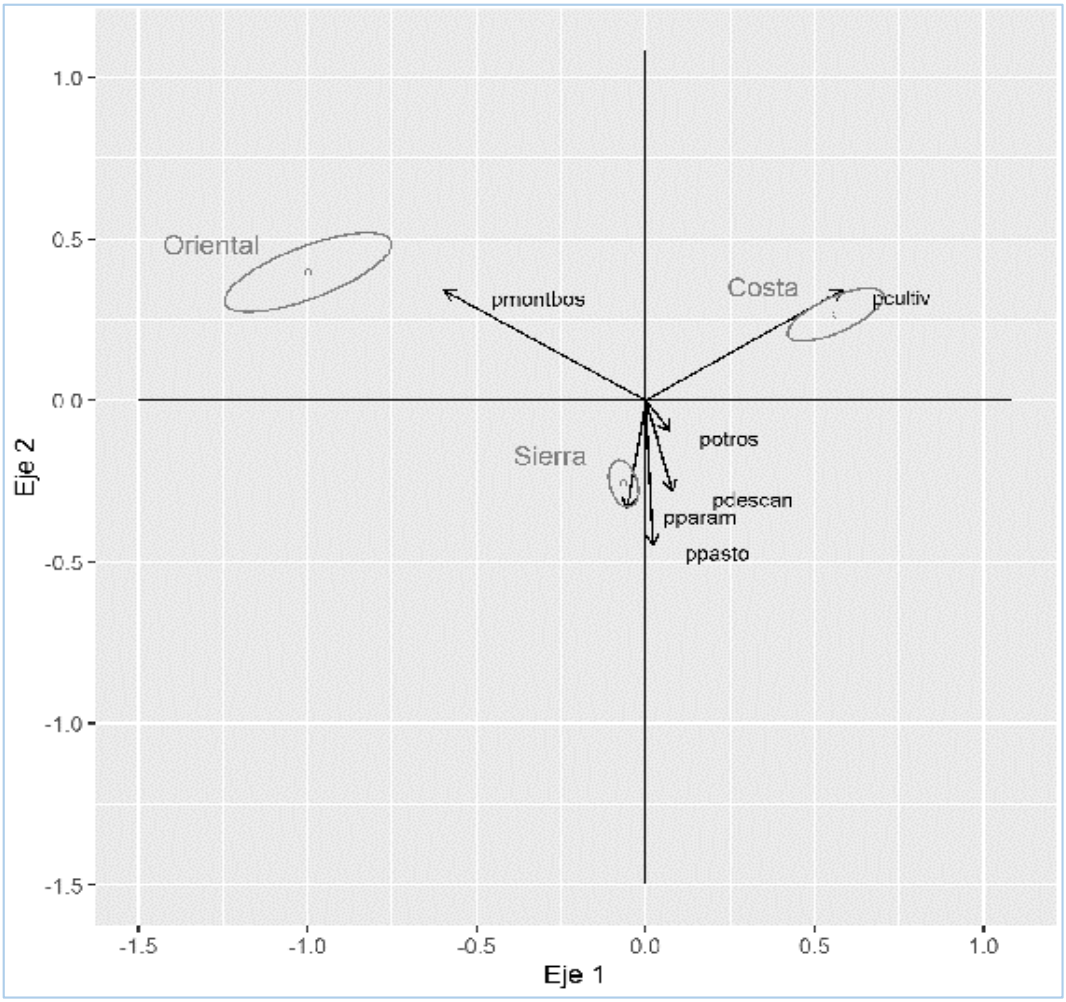

Figura 3. MANOVA-Biplot de grupos y variables.

Fuente: Elaboración propia a partir de la ESPAC2016.

En la **figura 3**, el primer eje es el que ofrece mayor discriminación entre grupos y explica el 75.13 % de la variabilidad (inercia), mientras que en conjunto con el segundo eje, se explica prácticamente el 100 %.

En la figura aparece la región oriental situada a la izquierda con altos valores para la variable pmontbos y la región Costa situada a la derecha con altos valores para la variable pcultiv.

La **tabla 4** contiene las calidades de representación de los grupos. Dichas calidades explican casi el 100 % de la variabilidad en los dos primeros ejes, por lo que la diferencia entre ellos puede interpretarse correctamente en las dos dimensiones que se presentan en la **figura 3**.

Tabla 4. Calidades de representación para los grupos.

Grupo	Eje 1 (%)	Eje 2 (%)	Total (%)
Sierra	6,03	93,97	100.00
Costa	81,74	18,26	100.00
Oriental	86,39	13,61	100.00

Fuente: Elaboración propia a partir de la ESPAC 2016.

La **tabla 5** contiene las calidades de representación de las variables. Aquellas que pueden ser correctamente interpretadas a partir de la **figura 3** son pcultiv y pmontbos, pues tienen una alta calidad de representación. Sobre las restantes no se logra una calidad de representación lo suficientemente elevada como para interpretarlas correctamente en dos dimensiones.

Tabla 5. Calidades de representación para las variables.

Variable	**Eje 1 (%)**	**Eje 2 (%)**	**Total (%)**
pcultiv	67,22	23,07	90,28
pdescan	1,12	14,33	15,45
ppasto	0,08	36,29	36,37
pparam	0,55	20,32	20,87
pmontbos	70,51	23,15	93,66
potros	0,91	1,56	2,47

Fuente: Elaboración propia a partir de la ESPAC 2016.

Con respecto a los supuestos de la técnica, en esta oportunidad se presenta una forma de representación gráfica (Biplot) que, aun cuando está relacionada con el MANOVA (que exige normalidad de los errores), utiliza argumentos de naturaleza puramente algebraica. Esto implica que la técnica es robusta frente a variaciones en la normalidad de los datos. En cuanto a la construcción de elipses de confianza, éstas si requieren del supuesto de normalidad, sin embargo, al tratarse de formas cuadráticas derivadas de centroides (o medias), con base en el Teorema del Límite Central puede suponerse el requisito satisfecho al menos de forma asintótica, con lo cual puede tenerse una confianza razonable en el resultado, en vista de que la muestra utilizada es suficientemente grande. En futuras

investigaciones, se propone estudiar la construcción de intervalos de confianza boostrap y no-paramétricos, que pueden también resultar apropiados en este contexto.

5.- Conclusiones

Se han desarrollado algunos aspectos inéditos de la teoría que soporta la técnica MANOVA-Biplot, principalmente en cuanto a la relación entre ambos conceptos (MANOVA y Biplot) y la elaboración de elipses de confianza, en lugar de las acostumbradas regiones de confianza circulares.

La propuesta de utilizar elipses de confianza en el Biplot muestra su valor en la aplicación a los datos del uso del suelo en la ESPAC 2016, ya que al tratarse de un conjunto voluminoso de datos, las regiones de confianza circulares resultan tan pequeñas que son inútiles para el análisis, debido a la disminución en la varianza de las medias a medida que aumenta el tamaño de la muestra.

Se puede apreciar cómo la región oriental está caracterizada por el uso de su suelo preponderantemente en montes, bosques naturales y artificiales, la región costera por cultivos permanentes, transitorios y barbechos y la región de la sierra por los restantes pastos, descansos, páramos y demás. Este resultado puede ser de ayuda considerable para la planificación productiva del uso del suelo en Ecuador, considerando las vocaciones de las regiones y pensando específicamente en las industrias, forestal, agrícola y pecuaria, respectivamente.

El análisis de los datos estudiados confirma las bondades y beneficios del MANOVA-Biplot respecto a otras técnicas estadísticas multivariantes, tales como el MANOVA, el Análisis de Variables Canónicas o el Análisis Factorial, sobre todo en cuanto a aspectos como la representación en el mismo gráfico de los centroides de los grupos, de las variables y de las regiones de confianza elípticas; y el uso de medidas de las calidades de representación que juegan un papel muy importante en la interpretación de los resultados, sobre todo por parte de investigadores de áreas aplicadas.

6.- Referencias

Amaro, I. R., Vicente-Villardón, J. L. & Galindo-Villardón, M. P. (2004). Manova-biplot para arreglos de tratamientos con dos factores basado en modelos lineales generales multivariantes. *Interciencia*, 29(1), 26-32.

Amaro, I. R., Vicente-Villardón, J. L. & Galindo-Villardón, M. P. (2008), Contribuciones al Manova-biplot: regiones de confianza alternativas. *Revista Investigación Operacional*, 29(3), 231-241.

Cuadras, C. M. (2014). *Nuevos Métodos de Análisis Multivariante*. Barcelona, España: CMC Editions.

Gabriel, K.R. (1972): Analysis of meteorological data by means of canonical decomposition and Biplots. Journal of Applied Meteorology, 11, 1071-1077.

Gabriel, K. R. (1995): MANOVA Biplots for two-way contingency tables. En: Recent Advances in Descriptive Multivariate Analysis. (W. KRZANOWSKI, ed.). Clarendon Press, Oxford, 227-268.

García-Talegón, J., Iñigo, A. C. & Vicente-Palacios, V. (2016), A laboratory simulation of desalting on calcareous building stone with wet sepiolite. *Environmental Earth Sciences*, 925(75), 1-15.

Gower, J. C. y Hand, D. J. (1996): Biplots. Chapman and Hall, London.

Iñigo, A. C., García-Talegón, J. & Vicente-Tavera, S. (2014), Canonical biplot statistical analysis to detect the magnitude of the effects of phosphates crystallization aging on the color in siliceous conglomerates. *Research and Application*, 39(1), 82-87.

Iñigo, A. C., García-Talegón, J. Vicente-Tavera, S., Casado-Marín, S. & Martín-Gonzales, S. (2017), Multivariate analyses of soluble salts responsible for pathologies in granites of the roman aqueduct of Segovia, Spain. *International Journal of Conservation Science*, 8(1), 59-66.

Iñigo, A. C., Vicente-Tavera, S. & Rives, V. (2004), Manova-biplot statistical analysis of the effect of artificial ageing (freezing/thawing) on the colour of treated granite stones. *Color Research and Applications*, 29(2), 115-120.

INEC (2017), Encuesta de Superficie y Producción Agropecuaria Continua ESPAC 2016: Informe Ejecutivo, Instituto Nacional de Estadística y Censos, Quito, Ecuador. Recuperado de: www.ecuadorencifras.gob.ec.

Mardia, K. V., Kent, J. T., Bibby, J.M. (1979). Multivariate Analysis. Academic press. London, RU.

Morrison, D. F. (1978). Multivariate Statistical Methods. McGraw-Hill. Londres, RU.

Nieto, A. B., Galindo, M. P., Leiva, V. & Vicente-Galindo, P. (2014). A methodology for biplots based on bootstrapping with R. *Revista Colombiana de Estadística*, 37(2), 367-397.

R Core Team. (2017). R: *A Language and Environment for Statistical Computing, R.* Vienna, Austria: Foundation for Statistical Computing. Recovered from: https://www.R-project.org/

Varas, M. J., Vicente-Tavera, S., Molina, E. & Vicente-Villardón, J. L. (2005). Role of canonical biplot method in the study of building stones: an example from Spanish monumental heritage. Environmetrics pp. 1-15.

Vicente-Villardón, J. L. (1992). Una alternativa a las técnicas factoriales basada en una generalización de los métodos Biplot. Tesis Doctoral. Universidad de Salamanca.

A. **Código R**

```
library(MASS)
library(ggplot2)
library(PostgreSQL)
source("./MyLib.R")
# Leyendo los datos
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, dbname = "ESPAC", host = "localhost",
                 port = 5432, user = "postgres", password = "")
dat <- dbGetQuery(con, "SELECT * FROM region_fin_p")
dat_a <- dbGetQuery(con, "SELECT*FROM region_fin_nn")
dbDisconnect(con)
dbUnloadDriver(drv)
head(dat)
dim(dat); dat_a
r <- sum(dat_a$n); s <- nrow(dat_a)
a <- matrix (0, r, s); k <- 1
for (j in 1: s) {
  for (i in 1: dat_a$n[j])  {
    a[k,j] <- 1
              k <- k + 1
  }
}
head(a); dim(a)
x <- as.matrix(dat[,2:7])
head(x)
nv <- as.vector(dat_a$n); nv
label1 <- as.vector(dat_a$region)
label2 <- as.vector(attributes(x)$dimnames[[2]])
n <- nrow(x); nuvar <- ncol(x); mg <- ncol(a)
cat("n", n, "nuvar", nuvar, "mg", mg,'\n')
x2 <- x; J <- matrix(1,n,n); I <- diag(1,n,n)
x1 <- (I-(1/n) *J) %*% x
head(x1)
#A continuación se hace la estandarización de la matriz x
X <- (I-(1/n) *J) %*% x
S <- diag((1/n) * (t(x) %*% x))
```

```
X <- x %*% solve(sqrt(diag(S)))

head(x)

#.................Cálculo de (inv(A'A))A'X).....

dim(x)

d <- t(a) %*% a

b <- solve(d) %*% t(a) %*% x

nt <- sum(nv); b; d

# ……Cálculo del vector x'…..

F <- t(nv) %*% b

f; nv

v <- as.vector(rep (1, mg)); v

#.........Cálculo de la matriz y

y <- b-v %*% f; y

va <- cov(x1); va

#…Cálculo de la matriz E.…matriz de sumas de cuadrados y productos 'dentro de'

E <- t(x) %*% x-t(b) %*% t(a) %*% a %*% b; e

b1 <- solve(d) %*% t(a) %*% x1

e1 <- t(x1) %*% x1-t(b1) %*% t(a) %*% a %*% b1

x1_cov <- cov(x1)

ho <- t(b1) %*% d %*% b1

eo <- e1

meo <- (1/(n-1)) * x1_cov

mo <- ho %*% solve(eo)

ev <- eigen(mo, symmetric=FALSE)

v <- ev$vectors; dj <- ev$values

norm(mo %*% v-diag(dj) %*% v)

dm <- solve(diag(sqrt(diag(x1_cov))))

ca <- dm %*% meo %*% v

ca <- ca*ca; ho; e1; meo; mo; dm

cat ('Las calidades de representación son:\n'); ca

#.......Descomposición en valores singulares de Y

e_inv <- solve(e)

e_inv_raiz <- raizMat(e_inv)

SVD <- svd(raizMat(d) %*% y %*% e_inv_raiz, nuvar, nuvar)

u <- SVD$u; l <- diag(SVD$d); m <- SVD$v

u; l; m

# …Proyección de los individuos
```

```
ind <- x %*% e_inv_raiz %*% m
#..... Cálculo de Inercias.
iner <- (100/sum(diag(l)^2)) * diag(l)^2
inercia <- cumsum(iner)
cat ('Inercia absorbida\n'); iner
cat ('Inercia acumulada\n'); inercia
cat ('P y Q\n');
# Marcadores P y Q del biplot
p <- raizMat(solve(d)) %*% u %*% l
q <- raizMat(e) %*% m
p; q
#***************programa KRAZNOSWKI
gama <- (mg-1) / (n-mg)
ese <- as.vector(c(nuvar-1, mg))
eses <- min(ese)
cat('eses', eses,'\n')
tita <- matrix(0, eses,1)
ll <- diag(l)
for (j in 1: eses) {
  tita[j] <- 1 + (1/(mg-1)) * ll[j]
}
cat('tita\n'); tita
#************ Cálculo de las varianzas
var <- matrix (0, mg,2)
for (i in 1:mg) {
  for (k in 1:2) {
    sumakra <- 0
            for (j in 1: eses) {
      if (j != k) {
        sumakra <- sumakra +
        ((tita[k]*(tita[j]-gama*tita[k]) /
        (gama*(tita[j]-tita[k])^2))*p[i, j]^2)
      }
            }
            var [i, k] <- 1/nv[i] + ((1/2)*p[i, k]^2+sumakra)/(n-mg)
  }
}
```

```
var
#************* Cálculo de las covarianzas
gamatita <- -((tita[1]*tita[2]*(1+gama)) / ((n-mg)*gama*(tita[1]-tita[2])^2))
mcov <- as.vector(c(rep (0,3)))
for (i in 1:3) {
  mcov[i] <- gamatita*p[i,1]*p[i,2]
}
mcov
#***************** Gráfica de las ELIPSES
#*********** Calidad de Representación **********
qs <- q^2
cat('q al cuadrado\n'); qs
qr <- matrix(0, nuvar, nuvar)
su <- as.vector(c(rep (0, nuvar)))
for (i in 1:nuvar) {
  for (k in 1:nuvar) {
    su[i] <- qs[i, k] + su[i]
  }
  for (j in 1: nuvar) {
    qr [i, j] <- qs [i, j]/su[i]
  }
}
cat ('La calidad de representación para las variables es:\n')
qr*100
sum (qr[1,1:3]); sum (qr[2,1:3]); sum (qr[3,1:3])
sum (qr[,1]); sum (qr[,2]); sum (qr[,3])
ps <- p^2
cat ('p al cuadrado\n'); ps
pqr <- matrix(0,3,3)
sup <- as.vector(c(rep (0,3)))
for (i in 1:3) {
  for (k in 1:3) {
    sup[i] <- ps[i, k] + sup[i]
  }
  for (j in 1:3) {
    pqr [i, j] <- ps[i, j] / sup[i]
  }
```

```
}
cat ('La calidad de representación para los grupos es:\n')
pqr * 100
neje <- 2
# .... Reescalado ....
escala <- 's'
if (escala == 's') {
  scp <- sum(p^2)
  scq <- sum(q^2)
  scp <- scp/mg
  scq <- scq/nuvar
  scf <- sqrt(sqrt(scq/scp))
  p <- p*scf # 1◦ y 2◦ columnas, coordenadas de los grupos G
  q <- q/scf; # 1◦ y 2◦ columnas, coordenadas de los vectores V
  ind <- ind*scf
}
# Representación Gráfica
p_graf <- as.data.frame(cbind(X=p[,1], Y=p[,2]))
q_graf <- as.data.frame(cbind(X=q[,1], Y=q[,2]))
p_graf; q_graf
ls <- max(max(p_graf), max(q_graf)) + 0.5
li <- min(min(p_graf), min(q_graf)) - 0.5
grf <- ggplot(NULL) +
  geom_point(aes (x = X, y = Y), data=p_graf,
    colour="red", shape=21, size=1) +
  geom_segment(aes(x = 0, y = 0, xend = q_graf$X, yend = q_graf$Y),
    arrow=arrow (length = unit (0.2, "cm"))) +
  geom_text(aes(x = X, y = Y, label=label1), data=p_graf,
    col="red", hjust=1.5, vjust=-1) +
  geom_text(aes(x = X, y = Y, label=label2), data=q_graf,
    hjust=-0.5, vjust=1, size=3) +
  scale_x_continuous("Eje 1", limits = c (li, ls)) +
  scale_y_continuous("Eje 2", limits = c (li, ls)) +
  geom_segment(aes(x = 0, y = li, xend = 0, yend = ls)) +
  geom_segment(aes(x = li, y = 0, xend = ls, yend = 0))
for (i in 1:3) {
  chi2 <- qchisq(0.05, 2, lower.tail = FALSE)
```

```
  MD <- matrix(c(var[i,1], mcov[i], mcov[i], var[i,2]), 2, 2)
  MD_Inv <- solve(MD)
  MD_f <- MD_Inv/chi2
  MD_ev <- eigen(MD_f, symmetric=F)
  angulo <- (180/pi) * acos(MD_ev$vectors[1,1])
  ejes <- sqrt(1/MD_ev$values)
  area <- pi*ejes[1]*ejes[2]
  cat("´ Angulo ", angulo, "\n")
  cat("Ejes ", ejes, "\n")
  cat("´ Area ", area, "\n")
  if (ejes[1] > ejes[2]) {
                eli <- elipseDat(ejeMay=ejes[1], ejeMen=ejes[2],
                        centro=c(p_graf$X[i], p_graf$Y[i]), a=angulo)
  } else {
                eli <- elipseDat(ejeMay=ejes[2], ejeMen=ejes[1],
                        centro=c(p_graf$X[i], p_graf$Y[i]), a=angulo)
              }
  grf <- grf +
    geom_path(data=eli, aes(x,y), col = 'green')
}
tiff ("Plot2.tif", width = 6, height = 6, units = 'in', res = 300)
grf
dev.off ()
```


ISNN 2588-0764