Kawashima Y. y Katagiri G. (1995). Fundamentals, overtones, and combinations in the Raman spectrum of graphite. *Phys. Rev. B*, 52, 10053–10059.

Lazzeri, M., Piscanec, S., Mauri, F., Ferrari A.C. y Robertson J. (2006). Phonon linewidths and electron-phonon coupling in graphite and nanotubes. *Phys. Rev. B*, 73, p. 155426.

Livneh, T., Haslett, T. L. y Moskovits, M. (2002). Distinguishing disorder-induced bands from allowed Raman bands in graphite. *Phys. Rev. B*, 66, p. 195110.

Machón, M., Reich, S. y Thomsen, C. (2002). Ab initio calculations of the optical properties of 4-Å-diameter single-walled nanotubes. *Phys. Rev. B*, 66, p. 155410.

Malard, L. M., Pimenta, M. A., Dresselhaus, G. y Dresselhaus, M. S. (2009). Raman spectroscopy in graphene. *Phys. Rep.*, 473, 51-87.

Maulzsch, J., Reich, S., Thompsen, C., Requardt, H. y Ordejón, P. (2004). Phonon Dispersion in Graphite. *Phys. Rev. Lett.*, 92, p. 075501.

Nemanich, R. J., Lucovsky, G. y Solin S. A., (1977). Infrared active optical vibrations of graphite. *Sol. Stat. Commun.* 23, 117-120.

Nemanich R. J. y Solin S. A. (1979). First- and second-order Raman scattering from finite-size crystals of graphite. *Phys. Rev. B*, 20, 392–401.

Nicklow R., Wakabayashi N. y Smith H.G. (1972). Lattice Dynamics of Pyrolytic Graphite, *Phys. Rev. B*, 5, 4951–4962.

Novoselov, K. S., Geim, A. K., Morozov, S. V., Jiang, D., Zhang, Y., Dubonos, S. V. Grigorieva, I. V. y Firsov, A. A. (2004). Electric Field Effect in Atomically Thin Carbon Films. *Science*, 306, 666-669.

Novoselov, K. S., Geim, A. K., Morozov, S. V., Jiang, D., Katsnelson, M. I., Grigorieva, I. V., Dubonos, S. V. y Firsov, A. A. (2005). Two-dimensional gas of massless Dirac fermions in grapheme. *Nature*, 438, 197-200.

Novoselov, K. S., Jiang, Z., Zhang, Y., Morozov, S. V., Stormer, H. L., Zeitler, U., Maan, J. C., Boebinger, G. S., Kim P. y Geim, A. K. (2007). Room-Temperature Quantum Hall Effect in Graphene. *Science*, 315, p. 1379.

Perdew, J. P., Burke, K. Ernzerhof M. (1996). Generalized Gradient Approximation Made Simple, *Phys. Rev. Lett.*, 77, 3865–3868.

Piscanec, S., Lazzeri, M., Mauri, F., Ferrari, A.C. y Robertson J. (2004). Kohn Anomalies and Electron-Phonon Interactions in Graphite. *Phys. Rev. Lett.*, 93, p. 185503.

Sánchez-Portal, D., Ordejón, P., Artacho E., Soler J. M. (1997). Density-functional method for very large systems with LCAO basis sets. *Int. J. Quantum Chem.*, 65, 453–461.

Shimada T., Sugai T., Fantini C., Souza M., Cançado L. G., Jorio A., Pimenta M. A., Saito

R., Grüneis A., Dresselhaus G., Dresselhaus M. S., Ohno Y., Mizutani T. y Shinohara H. (2005). Origin of the 2450 $cm^{-1}$ Raman bands in HOPG, single-wall and double-wall carbon nanotubes. *Carbon*, 43, 1049-1054.

Soler, J. M., Artacho, E., Gale, J. D., Garcia, A., Junquera, J., Ordejón, P. y Sánchez-Portal, D. (2002). The SIESTA method for ab initio order-N materials simulation. *J. Phys.: Condens. Matter*, 14, 2745-2779.

Stephanie R. y Thomsen C. (2004). Raman spectroscopy of graphite. *Phil. Trans. R. Soc. Lond. A*, 362, 2271-2288.

Tan, P. H., Hu, C. Y., Dong, J., Shen, W. C. y Zhang B.F. (2001). Polarization properties, high-order Raman spectra, and frequency asymmetry between Stokes and anti-Stokes scattering of Raman modes in a graphite whisker. *Phys. Rev. B*, 64, p. 214301.

Thomsen, C. y Reich, S. (2000). Double Resonant Raman Scattering in Graphite. *Phys. Rev. Lett.*, 85, 5214–5217.

Troullier, N. y Martins, J. L. (1991). Efficient pseudopotentials for plane-wave calculations, *Phys. Rev. B*, 43, p. 1993.

Tuinstra, F. y Koening, J. L. (1970). Raman Spectrum of Graphite. *J. Chem. Phys.*, 53, 1126-1130.

Vidano, R. P., Fischbach, D. B., Willis, L. J. y Loehr, T. M., (1981). Observation of Raman band shifting with excitation wavelength for carbons and graphites. *Sol. Stat. Commun.*, 39, 341-344.

Wang, Y. Y., Ni, Z. H., Yu, T., Shen, Z. X., Wang, H. M., Wu, Y. H., Chen, W. y Wee A. T. S. (2008). Raman Studies of Monolayer Graphene: The Substrate Effect. *J. Phys. Chem. C*, 112, 10637-10640.

Wilson, M. (2006). Scanning Tunneling Microscope Measures the Spin-Excitation Spectrum of Atomic-Scale Magnets. *Physics Today*, 59, 13-14.

Wirtz, L. y Rubio, A. (2004). The phonon dispersion of graphite revisited. *Sol. Stat. Commun.*, 131, 141-152.



| Referencia | Fecha de recepción | Fecha de aprobación |
|---|---|---|
| Marquina, J.; Power, Ch. y González, J.<br>Espectroscopia Raman del grafeno monocapa y el grafito: acoplamiento electrón fonón y efectos no adiabáticos.<br>Revista *Tumbaga* (2010), 5, 183-194 | Día/mes/año<br>18/09/2010 | Día/mes/año<br>22/10/2010 |

# Las bases de Gröbner en el estudio de los polinomios simétricos

Cifuentes, V.[I]; Patiño, B.; Pérez, H.[II]

**Resumen.** En este artículo presentamos dos algoritmos, el primero permite escribir un polinomio simétrico $f$ en $k[x_1, \cdots, x_n]$, con $k$ un cuerpo, en términos de las funciones simétricas elementales; el segundo, determina si un polinomio $f$ en $k[x_1, \cdots, x_n]$, con $k$ un cuerpo, es simétrico, y si este es el caso, cómo escribirlo en términos de las funciones simétricas elementales. Además, probamos de manera detallada cómo se obtiene una base de Gröbner $G$ en el caso particular cuando se considera el orden $lex$ sobre los términos, herramienta necesaria para presentar el segundo algortimo. Adicionalmente, mostramos una pequeña aplicación de los polinomios simétricos en el cálculo del anillo de invariantes de un grupo finito de matrices dado. Ilustramos los resultados con variados ejemplos.

**Palabras clave:** Polinomios simétricos, polinomios simétricos elementales, bases de Gröbner, anillo de invariantes.

**Abstract.** In this article we present two algorithms, the first one allows to write a polynomial $f \in k[x_1, \cdots, x_n]$, with $k$ a field, in terms of the symmetrical elementary functions, the second one determines if a polynomial $f \in k[x_1, \cdots, x_n]$, with $k$ a field, is symmetrical, and if this one is the case, how to write it in terms of the symmetrical elementary functions. As complement, we show in a detailed way how Gröbner's base is obtained in the particular case when the order is considered to be $lex$, necessary tool to present the second algorithm. Finally we present a small application of the symmetrical polynomials in the calculation of the rings of invariants of a finite matrix groups. We ilustrate the results with several examples.

**Keywords:** Symmetric polinomials, elementary symmetric functions, Gröbner bases, rings of invariants.

I Docente escuela de matemáticas, Universidad Pedagógica y Tecnológica de Colombia, Tunja, Colombia. veciva@yahoo.com.

II Estudiantes licenciatura en matemáticas, Universidad Pedagógica y Tecnológica de Colombia, Tunja, Colombia. azulcielo22@hotmail.com, hperez042000@yahoo.com

**Clasificación de materias (AMS):** 53A35,58A05,53A35.

# 1. INTRODUCCIÓN

Durante las últimas décadas se han logrado avances significativos en el desarrollo de métodos algorítmicos en matemáticas, tanto para realizar cálculos que manualmente son bastante extensos o que a veces pueden tornarse difíciles, como para demostrar proposiciones y teoremas. En álgebra conmutativa la teoría y métodos de las bases de Gröbner, permiten realizar cálculos efectivos en $k[x_1, \cdots, x_n]$, el anillo de polinomios en $n$ indeterminadas con coeficientes en un cuerpo $k$. Los paquetes computacionales como Singular, CoCoa, Maple, entre otros, cuentan con una librería que permite realizar los cálculos anteriormente mencionados usando dicha técnica.

En el estudio de los polinomios simétricos, las bases de Gröbner nos permiten determinar si un polinomio es simétrico y si este es el caso, cómo escribirlo en términos de los polinomios simétricos elementales. Este procedimiento es constructivo y por tanto podemos presentar un algoritmo, el cual es una motivación para el estudio de la teoría de invariantes desde el punto de vista computacional, es decir, de calcular de manera explícita el anillo de invariantes de un grupo finito.

# 2. PRELIMINARES

## 2.1 Orden de términos

**Definición 2.1.** *Un producto de potencias en $A = k[x_1, \cdots, x_n]$ es una expresión de la forma $X^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, donde $\alpha := (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n$. El conjunto de todos los productos de potencias será denotado por*

$$\mathbb{T}^n = \{x_1^{\alpha_1} \cdots x_n^{\alpha_n} \mid \alpha_i \in \mathbb{N}, i = 1, \cdots, n\}$$

**Definición 2.2.** *Un polinomio $f \neq 0 \in k[x_1, \cdots, x_n]$ es una suma finita de términos de la forma $a_i x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, con $a_i \neq 0 \in k$, es decir,*

$$f = a_1 X^{\alpha_1} + \cdots + a_t X^{\alpha_t}$$

*donde, $X^{\alpha_1} > X^{\alpha_2} > \cdots > X^{\alpha_t}$ y $X^{\alpha_i} \in \mathbb{T}^n$. En este caso*

- *$lp(f) = X^{\alpha_1}$, es el producto de potencias principal de $f$.*
- *$lc(f) = a_1$, es el coeficiente principal de $f$.*

- $lt(f) = a_1 X^{\alpha_1}$, *es el término principal de* $f$.

**Definición 2.3.** *Un orden de términos es un orden total* $<$ *que satisface las siguientes dos condiciones:*

(i) $1 < X^\alpha$ *para todo* $X^\alpha \in \mathbb{T}^n$, $X^\alpha \neq 1$.

(ii) *Si* $X^\alpha < X^\beta$, *entonces* $X^\alpha X^\gamma < X^\beta X^\gamma$ *para todo* $X^\gamma \in \mathbb{T}^n$.

Existen diferentes ordenes, se presentan a continuación los tres más conocidos en la literatura, éstos son usados en los paquetes computacionales existentes como CoCoa, Maple, Singular.

**Definición 2.4.** *Sean* $X^\alpha < X^\beta$ *productos de potencias, se definen los siguientes ordenes sobre* $\mathbb{T}^n$ *con* $x_1 > x_2 > \cdots > x_n$.

(i) *El orden lex (lexicográfico)*

$$X^\alpha < X^\beta \Leftrightarrow \begin{cases} \text{la primera coordenada } \alpha_i \text{ y } \beta_i \text{ en } \alpha \text{ y } \beta \text{ de izquierda a} \\ \text{derecha, las cuales son diferentes satisfacen } \alpha_i < \beta_i \end{cases}$$

*donde* $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n), \beta = (\beta_1, \beta_2, \cdots, \beta_n) \in \mathbb{N}^n$

(ii) *El orden deglex (lexicográfico de grado)*

$$X^\alpha < X^\beta \Leftrightarrow \begin{cases} \sum_{i=1}^n \alpha_i < \sum_{i=1}^n \beta_i \\ o \\ \sum_{i=1}^n \alpha_i = \sum_{i=1}^n \beta_i \quad y\ X^\alpha < X^\beta \text{ con respecto a lex} \\ \text{con } x_1 > x_2 > \cdots > x_n. \end{cases}$$

*donde* $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n) \in \mathbb{N}^n, \beta = (\beta_1, \beta_2, \cdots, \beta_n) \in \mathbb{N}^n$

(iii) *El orden degrevlex (lexicográfico de grado reverso)*

$$X^\alpha < X^\beta \Leftrightarrow \begin{cases} \sum_{i=1}^n \alpha_i < \sum_{i=1}^n \beta_i \\ o \\ \sum_{i=1}^n \alpha_i = \sum_{i=1}^n \beta_i \text{ y la primera coordenada de } \alpha_i \text{ y } \beta_i \text{ en} \\ \alpha \text{ y } \beta \text{ desde la derecha, las cuales son diferentes satisfacen} \\ \alpha_i > \beta_i. \end{cases}$$

*donde* $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n) \in \mathbb{N}^n, \beta = (\beta_1, \beta_2, \cdots, \beta_n) \in \mathbb{N}^n$

## 2.2 Nociones básicas de bases de Gröbner

En esta sección presentamos las nociones básicas de la teoría de las bases de Gröbner, que serán necesarias en las secciones posteriores. Un estudio detallado de esta teoría se puede hacer siguiendo el libro de Adams y Loustaunau (veáse [1]).

**Definición 2.5.** *Un conjunto de polinomios distinto de cero* $G = \{g_1, \cdots, g_t\}$ *contenido en un ideal* $I :=< f_1, \cdots, f_k >$*, es una base de Gröbner para* $I$ *si, y sólo si, para todo* $f \in I, f \neq 0$*, existe* $i \in \{1, \cdots, t\}$ *tal que* $lp(g_i)$ *divide a* $lp(f)$*.*

**Definición 2.6.** *Sean* $0 \neq f, g \in k[x_1, \cdots, x_n]$*. Se define el mínimo común múltiplo de* $f$ *y* $g$*, denotado por* $lcm(f, g)$*, al polinomio* $l$ *tal que:*

(i) $f, g$ *dividen a* $l$*.*

(ii) *Si* $f, g$ *dividen a un polinomio* $h$*, entonces* $l$ *divide a* $h$*.*

(iii) $lc(l) = lc(f)lc(g)$*.*

**Definición 2.7.** *Sean* $0 \neq f, g \in k[x_1, \cdots, x_n]$*. Sea* $L = lcm(lp(f), lp(g))$*. El polinomio*

$$S(f, g) = \frac{L}{lt(f)} f - \frac{L}{lt(g)} g$$

*se denomina el* $S$*- polinomio de* $f$ *y* $g$*.*

**Teorema 2.8.** *Sea* $G = \{g_1, \cdots, g_t\}$ *un conjunto de polinomios no nulos en* $k[x_1, \cdots, x_n]$*. Entonces* $G$ *es una base de Gröbner para el ideal* $I =< g_1, \cdots, g_t >$ *si, y sólo si, para todo* $i \neq j$

$$S(g_i, g_j) \xrightarrow{G}_+ 0,$$

**Proposición 2.9.** *Sea* $G \subset k[x_1, \cdots, x_n]$ *un conjunto finito y sean* $f, g \in G$ *tales que*

$$lcm(lp(f), lp(g)) = lp(f) \cdot lp(g)$$

*es decir que los monomios principales de* $f$ *y* $g$ *son primos relativos. Entonces*

$$S(g_i, g_j) \xrightarrow{G}_+ 0,$$

## 3. POLINOMIOS SIMÉTRICOS

Cuando se estudian las raíces de un polinomio surgen de manera natural polinomios simétricos, los cuales reciben el nombre de funciones simétricas elementales. Éstas juegan un papel fundamental ya que cualquier polinomio simétrico en $k[x_1, \cdots, x_n]$ puede ser escrito en términos de dichas funciones. En esta sección mostraremos un algoritmo que permite hacer dicho procedimiento.

**Definición 3.1.** *Un polinomio $f \in k[x_1, \cdots, x_n]$ es simétrico si*

$$f(x_{i_1}, \cdots, x_{i_n}) = f(x_1, \cdots, x_n)$$

*para todas las posibles permutaciones $x_{i_1}, \cdots, x_{i_n}$ de las variables $x_1, \cdots, x_n$.*

**Ejemplo 3.2.** *Sea $A = \mathbb{Q}[x, y, z]$, entonces el polinomio $f(x, y, z) = x^3 + x^2y^2z^2 + y^3 + z^3$ es simétrico ya que*

$$f(x, y, z) = f(x, z, y) = f(y, x, z) = f(y, z, x) = f(z, x, y) = f(z, y, x).$$

**Definición 3.3.** *Dadas las variables $x_1, \cdots, x_n$, se define $\sigma_1, \cdots, \sigma_n \in k[x_1, \cdots, x_n]$ de la siguiente manera:*

$$\begin{aligned}
\sigma_1 &= x_1 + \cdots + x_n \\
&\vdots \\
\sigma_i &= \sum_{j_1 < j_2 < \cdots < j_i} x_{j1} x_{j2} x_{j3} \cdots x_{ji} \\
&\vdots \\
\sigma_n &= x_1 x_2 x_3 \cdots x_n
\end{aligned}$$

**Proposición 3.4.** *Si $x_1, \cdots, x_n$ son las raíces de un polinomio $f(x)$, entonces $f(x)$ puede ser expresado usando las funciones $\sigma_1, \cdots, \sigma_n$ de la siguiente manera*

$$f(x) = x^n - \sigma_1 x^{n-1} + \sigma_2 x^{n-2} + \cdots + (-1)^{n-1}\sigma_{n-1}x + (-1)^n\sigma_n \qquad (1)$$

**Proposición 3.5.** *Los polinomios $\sigma_1, \cdots, \sigma_n$ en la definición 3.3 son simétricos y se denominan las funciones simétrica elementales.*

*Demostración.* Usando la proposición anterior, ya que $x_1, \cdots, x_n$ son las raíces de un polinomio $f(x)$, podemos escribir $f$ de la siguiente manera

$$f(x) = (x - x_1)(x - x_2) \cdots (x - x_n) \qquad (2)$$

luego, al realizar cualquier permutación $x_{i_1}, \cdots, x_{i_n}$ de las variables $x_1, \cdots, x_n$, se obtiene el mismo polinomio salvo por el orden de los factores, así, los coeficientes en (3.1) son funciones simétricas. □

**Teorema 3.6.** *Teorema fundamental de polinomios simétricos. Cualquier polinomio simétrico en $k[x_1, \cdots, x_n]$ pueden ser escrito de manera única como un polinomio en las funciones simétricas elementales $\sigma_1, \cdots, \sigma_n$.*

*Demostración.* La prueba puede ser consultada en [2]. □

La demostración del teorema anterior nos permite presentar un algoritmo para escribir cualquier polinomio simétrico $f \in k[x_1, \cdots, x_n]$ en términos de los polinomios simétricos elementales.

**Algoritmo para polinomios simétricos**

**ENTRADA:** $f \neq 0 \in k[x_1, \cdots, x_n]$ polinomio simétrico
$\sigma_1, \cdots, \sigma_n$ las funciones simétricas elementales.
**SALIDA:** $a_1, a_2, \cdots, a_s, h_1, h_2, \cdots, h_s$ tal que $f = a_1h_1 + \cdots + a_sh_s$
**INICIO:** $a_1 := 0, a_2 := 0, \cdots, a_s := 0, h_1 := 0, h_2 := 0, \cdots, h_s := 0$
$p := f$

**MIENTRAS** $p \neq 0$ **HAGA**

Calcule $lt(p) = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$

$h_i := h_i + \sigma_1^{\alpha_1 - \alpha_2} \sigma_2^{\alpha_2 - \alpha_3} \cdots \sigma_{n-1}^{\alpha_{n-1} - \alpha_n} \sigma_n^{\alpha_n}$.
$a_i := a_i + \frac{lt(p)}{lt(h_i)}$

$p := p - a_i h_i$

$f = a_1h_1 + a_2h_2 + \cdots + a_sh_s$

**Ejemplo 3.7.** *Considere el polinomio*

$$f = (x^2 + y^2)(x^2 + z^2)(y^2 + z^2) \in k[x, y, z]$$

*escriba $f$ como un polinomio en las funciones simétricas elementales $\sigma_1, \sigma_2, \sigma_3$, donde, $x > y > z$ y se considera el orden lex.*

$$f(x,y,z) = x^4y^2 + x^4z^2 + x^2y^4 + 2x^2y^2z^2 + x^2z^4 + y^4z^2 + y^2z^4$$

*Primer paso a través del mientras:*
$lt(p) = x^4y^2$;

$$\begin{aligned} h_1 &= \sigma_1^2\sigma_2^2 \\ &= x^4y^2 + 2x^4yz + x^4z^2 + 2x^3y^3 + 8x^3y^2z + x^3yz^2 + 2x^3z^3 + y^2y^4 + 8x^2y^3z + 15x^2y^2z^2 \\ &\quad + 8x^2yz^3 + x^2z^4 + 2xy^4z + 8xy^3z^2 + 8xy^2z^3 + 2xyz^4 + y^4z^2 + 2y^3z^3 + y^2z^4 \end{aligned}$$

$lt(h_1) = x^4y^2; a = \frac{x^4y^2}{x^4y^2} = 1$

$$\begin{aligned} p &= f - a_1h_1 = f - \sigma_1^2\sigma_2^2 \\ &= -2x^4yz - 2x^3y^3 - 8x^3y^2z - 8x^3yz^2 - 2x^3z^3 - 8x^2y^3z - 13x^2y^2z^2 - 8x^2yz^3 \\ &\quad - 2xy^4z - 8xy^3z^2 - 8xy^2z^3 - 2xyz^4 - 2y^3z^3 \end{aligned}$$

*Segundo paso a través del mientras:*
$lt(p) = -2x^4yz$;

$$\begin{aligned} h_2 &= \sigma_1^3\sigma_3^1 \\ &= x^4yz + 3x^3y^2z + 3x^3yz^2 + 3x^2y^3z + 6x^2y^2z^2 + 3x^2yz^3 + xy^4z + 3xy^3z^2 + 3xy^2z^3 + xyz^4 \end{aligned}$$

$lt(h_2) = x^4yz; a_2 = \frac{-2x^4yz}{x^4yz} = -2$

$$\begin{aligned} p &= p - a_2h_2 = f - \sigma_1^2\sigma_2^2 + 2\sigma_1^3\sigma_3^1 \\ &= -2x^3y^3 - 2x^3y^2z - 2x^3yz^2 - 2x^3z^3 - 2x^2y^3z - x^2y^2z^2 - 2x^2yz^3 - 2xy^3z^2 - 2xy^2z^3 - 2y^3z^3 \end{aligned}$$

*Tercer paso a través del mientras:*
$lt(p) = -2x^3y^3$;

$$\begin{aligned} h_3 &= \sigma_2^3 \\ &= x^3y^3 + 3x^3y^2z + 3x^3yz^2 + x^3z^3 + 3x^2y^3z + 6x^2y^2z^2 + 3x^2yz^3 + 3xy^3z^2 \quad + 3xy^2z^3 + y^3z^3 \end{aligned}$$

$lt(h_3) = x^3y^3; a_3 = \frac{-2x^3y^3}{x^3y^3} = -2$

$$\begin{aligned} p &= p - a_3h_3 = f - \sigma_1^2\sigma_2^2 + 2\sigma_1^3\sigma_3^1 + 2\sigma_2^3 \\ &= 4x^3y^2z + 4x^3yz^2 + 4x^2y^3z + 11x^2y^2z^2 + 4x^2yz^3 + 4xy^3z^2 + 4xy^2z^3 \end{aligned}$$

*Cuarto paso a través del mientras:*
$lt(p) = 4x^3y^2z$;

$$\begin{aligned} h_4 &= \sigma_1\sigma_2\sigma_3 \\ &= x^3y^2z + x^3yz^2 + x^2y^3z + 3x^2y^2z^2 + x^2yz^3 + xy^3z^2 + xy^2z^3 \end{aligned}$$

$lt(h_4) = x^3y^2z, a_4 = \frac{4x^3y^2z}{x^3y^2z} = 4$

$p = p - a_4h_4 = f - \sigma_1^2\sigma_2^2 + 2\sigma_1^3\sigma_3^1 + 2\sigma_2^3 - 4\sigma_1\sigma_2\sigma_3 = x^2y^2z^2$

*Quinto paso a través del mientras:*
$lt(p) = x^2y^2z^2$

$$h_5 = \sigma_3^2 = x^2y^2z^2$$

$lt(h_5) = x^2y^2z^2, a_5 = \frac{x^2y^2z^2}{x^2y^2z^2} = 1$

$p = f - a_5h_5 = f - \sigma_1^2\sigma_2^2 + 2\sigma_1^3\sigma_3^1 + 2\sigma_2^3 - 4\sigma_1\sigma_2\sigma_3 - \sigma_3^2 = 0$
*Ya que* $p = 0$ *el ciclo mientras termina y se obtiene,*

$$f = \sigma_1^2\sigma_2^2 - 2\sigma_1^3\sigma_3^1 - 2\sigma_2^3 + 4\sigma_1\sigma_2\sigma_3 + \sigma_3^2$$

## 4. BASES DE GRÖBNER Y POLINOMIOS SIMÉTRICOS

Las bases de Gröbner son una herramienta útil en el estudio de los polinomios simétricos, ya que permiten determinar si un polinomio en $k[x_1, \cdots, x_n]$ es simétrico, y en caso afirmativo expresar $f$ en términos de los polinomios simétricos elementales.

**Proposición 4.1.** *En el anillo* $k[x_1, \cdots, x_n, y_1, \cdots, y_n]$ *se fija un orden en los monomios de tal manera que un monomio que contenga una de las variables* $x_1, \cdots, x_n$ *es mayor que cualquier monomio en* $k[y_1, \cdots, y_n]$. *Sea* $G$ *una base de Gröbner del ideal* $\langle \sigma_1 - y_1, \cdots, \sigma_n - y_n \rangle \subset k[x_1, \cdots, x_n, y_1, \cdots, y_n]$. *Dado* $f \in k[x_1, \cdots, x_n]$ *y* $g = \bar{f}^G$ *el residuo de* $f$ *al dividirlo por* $G$. *Entonces:*

(i) $f$ *es simétrico si, y sólo si,* $g \in k[y_1, \cdots, y_n]$

(ii) *Si* $f$ *es simétrico, entonces* $f = g(\sigma_1, \cdots, \sigma_n)$ *es la única expresión de* $f$ *como un polinomio en las funciones simétricas elementales* $\sigma_1, \cdots, \sigma_n$

La proposición anterior muestra la necesidad de conocer métodos para calcular una base de Gröbner para $\langle \sigma_1 - y_1, \cdots, \sigma_n - y_n \rangle$. Una manera de hacerlo es usando el algoritmo de Buchberger.[1] Sin embargo, cuando se usa el orden lex, hay un método bastante sencillo para calcular una base para dicho ideal, el cual mostramos en detalle a continuación.

**Proposición 4.2.** *Fijado el orden lex sobre* $k[x_1, \cdots, x_n, y_1, \cdots, y_n]$ *con* $x_1 > \cdots > x_n > y_1 > \cdots > y_n$. *Entonces los polinomios*

[1] Para más detalles veáse [1]

$$g_k = h_k(x_k, \cdots, x_n) + \sum_{i=1}^{k}(-1)^i h_{k-i}(x_k, \cdots, x_n)y_i,\ k = 1, \cdots, n,$$

*forman una base de Gröbner para el ideal* $\langle \sigma_1 - y_1, \cdots, \sigma_n - y_n \rangle$*, donde*

$$h_i(x_1, \cdots, x_s) = \sum_{|\alpha|=i}(X^\alpha)$$

*es la suma de todos los monomios de grado total* $i$ *en* $x_1, \cdots, x_s$*.*

*Demostración.* En primer lugar se debe probar que el conjunto de los $g_k, k = 1 \cdots, n$ son un subconjunto del ideal $\langle \sigma_1 - y_1, \cdots, \sigma_n - y_n \rangle$.
*Paso 1.* Probar que

$$0 = h_k(x_k, \cdots, x_n) + \sum_{i=1}^{k} h_{k-i}(x_k, \cdots, x_n)\sigma_i$$

*Paso 1.1* Probar que $0 = \sum_{i=0}^{k}(-1)^i h_{k-i}(x_1, \cdots, x_n)\sigma_i(x_1, \cdots, x_n)$
En la demostración denotaremos $x := x_1, \cdots, x_n$. Si $X^\alpha = x_{j_1}^{\alpha_1} x_{j_2}^{\alpha_2} \cdots x_{j_a}^{\alpha_a}$ es un monomio que aparece en $h_{k-i}(x)\sigma_i(x)$, donde $a$ denota el número de variables que aparecen en $X^\alpha$ entonces se debe tener que $i \leq a$, en efecto, ya que $\sigma_i$ es la suma de todos los monomios que son productos de $i$ distintas variables, entonces cada término que aparece en el producto $h_{k-i}(x)\sigma_i(x)$ deben involucrar como mínimo las $i$ variables que aparecen en $\sigma_i(x)$. Así, el número de variables que aparecen en $X^\alpha$ es mayor o igual a $i$, es decir, $a \geq i$.
Ahora, ya que $i \leq a$ entonces determinar todos los monomios que involucren $i$ variables de las $a$ variables dadas en $X^\alpha$ , se reduce a resolver un problema de combinatoria ya que al usar cualquier orden de términos hay solamente una forma de escribir cada monomio. Por cada combinación $C_{a,i}$ se obtienen $i!$ permutaciones, luego $C_{a,i} \times r! = P_{a,i} = \frac{a!}{(a-i)!}$, donde $P_{a,i}$ denota el número de permutaciones, es decir $C_{a,i} = \frac{a!}{(a-i)i!} = \binom{a}{i}$, las cuales según la definición de $\sigma_i$, son monomios que aparecen allí, por tanto, hay $\binom{a}{i}$ términos de $\sigma_i(x)$ que aparecen en $X^\alpha$. Ya que $h_{k-i}(x)$ es la suma de todos los monomios de grado total $k - i$ en $x_1, \cdots, x_n$, los cuales tiene coeficiente 1, y existen $\binom{a}{i}$ términos de $\sigma_i$ que involucran variables de $X^\alpha$, entonces $X^\alpha$ aparecerá $\binom{a}{i}$ veces en $h_{k-i}(x)\sigma_i(x)$, así, el coeficiente de $X^\alpha$ en $\sum_{i=0}^{k}(-1)^i h_{k-i}(x)\sigma_i(x)$ es $\sum_{i=0}^{a}(-1)^i\binom{a}{i}$. Aplicando el teorema del binomio se obtiene

$$0 = (-1+1)^a = \sum_{i=0}^{a}\binom{a}{i}(-1)^i(1)^{a-i} = \sum_{i=0}^{a}\binom{a}{i}(-1)^i$$

es decir, el coeficiente con que aparece $X^\alpha$ en $h_{k-i}(x)\sigma_i(x)$ es cero, por tanto,

$$0 = \sum_{i=0}^{k} (-1)^i h_{k-i}(x_1, \cdots, x_n)\sigma_i(x_1, \cdots, x_n)$$

*Paso 1.2* Probar que

$$0 = h_k(x_k, \cdots, x_n) + \sum_{i=1}^{k} (-1)^i h_{k-i}(x_k, \cdots, x_n)\sigma_i(x_1, \cdots, x_n), \tag{3}$$

lo cual es equivalente a probar que

$$0 = \sum_{i=0}^{k} (-1)^i h_{k-i}(x_1, \cdots, x_n)\sigma_i(x_1, \cdots, x_n).$$

Para usar la identidad probada en el paso 1.1, se deben separar las variables $x_1, \cdots, x_{k-1}$, para esto, sea $A = \{1, 2, \cdots, k-1\}$, el conjunto formado por los subíndices de las variables $x_i, i = 1, \cdots, k-1$, $S \subset A$, $X^S$ el producto de las correspondientes variables según los subíndices involucrados en $S$, $|S|$ el número de elementos en $S$ y sea

$$H = \{S/S \subset A\} \qquad = \{\emptyset, \{1\}, \{2\}, \cdots, \{k-1\}, \{1,2\}, \{1,3\}, \cdots, \{1,2,\cdots,k-2\}\}$$

Mostraremos que $\sum_{S \in H} X^s \sigma_{i-|S|}(x_k, \cdots, x_n) = \sigma_i(x_1, \cdots, x_n)$

Para mayor facilidad tomaremos $y = x_k, \cdots, x_n$ y $x = x_1, \cdots, x_{k-1}$

$$\begin{aligned}
\sum_{S \in H} X^s \sigma_{i-|S|}(y) = \sigma_i(y) + x_1\sigma_{i-1}(y) + x_2\sigma_{i-1}(y) + \cdots + x_{k-1}\sigma_{i-1}(y) + \\
x_1x_2\sigma_{i-2}(y) + \cdots + x_1x_{k-1}\sigma_{i-2}(y) + \cdots + x_{k-2}x_{k-1} \\
\sigma_{i-2}(y) + \cdots + x_1x_2\cdots x_{k-1}\sigma_{i-\{k-1\}}(y) \\
= \sigma_i(y) + \sigma_{i-1}(y)(x_1 + x_2 + \cdots + x_{k-1}) + \sigma_{i-2}(y)(x_1x_2 \\
+ \cdots \quad + x_{k-2}x_{k-1}) + \cdots + \sigma_{i-k+1}(y)(x_1x_2\cdots x_{k-1}) \\
= \sigma_i(y) + \sigma_{i-1}(y)\sigma_1(x) + \sigma_{i-2}(y)\sigma_2(x) + \sigma_{i-3}(y)\sigma_3(x)\cdots \\
+ \sigma_{i-(k-2)}(y)\sigma_{k-2}(x)
\end{aligned}$$

Ya que cada producto $\sigma_{i-j}(y)\sigma_j(x), j = 0, \cdots, k-2$, es la suma de todos los monomios de $i$ variables distintas de las variables involucradas en cada uno de ellos, se obtiene la identidad deseada.

$$\begin{aligned}
\sum_{i=0}^{k} (-1)^i h_{k-i}(y)\sigma_i(x) &= \sum_{i=0}^{k} (-1)^i h_{k-i}(y) \sum_{S} \in HX^S\sigma_{i-|S|}(y) \\
&= h_k(y) \sum_{S \in H} X^S\sigma_{0-|S|}(y) - h_{k-1}(y) \sum_{S \in H} X^S\sigma_{1-|S|}(y) + h_{k-2}(y) \\
&\quad \sum_{S \in H} X^S\sigma_{2-|S|}(y) + \cdots + (-1)^k h_0(y) \sum_{S \in H} X^S\sigma_{k-|S|}(y) \\
&= \sum_{S \in H} X^S[h_k(y)\sigma_{0-|S|}(y) - h_{k-1}(y)\sigma_{1-|S|}(y) + h_{k-2}(y)\sigma_{2-|S|}(y) \\
&\quad + \cdots + (-1)^k h_0(y)\sigma_{k-|S|}(y)]
\end{aligned}$$

$$= \sum_{S\in H} X^S[\sum_{i=|S|}^{k}(-1)^i h_{k-i}(y)\sigma_{k-|S|}(y)]$$

donde la suma $\sum_{i=|S|}^{k}(-1)^i h_{k-i}(y))\sigma_{k-|S|}(y) = 0$, para cada $S \in H$. En efecto, haciendo la sustitución $j = i - |s|$, se obtiene,

$$\sum_{i=|S|}^{k}(-1)^i h_{k-i}(y))\sigma_{k-|S|}(y) = (-1)^{|s|}\sum_{j=0}^{k-|s|}(-1)^j h_{(k-|s|)-j}(y)\sigma_j(y) = 0$$

usando la identidad del paso 1.1. Así,

$$\sum_{i=0}^{k}(-1)^i h_{k-i}(y)\sigma_i(x) = \sum_{S\in H} X^S(\sum_{i=|S|}^{k}(-1)^i h_{k-i}(y)\sigma_{k-|S|}(y)) = \sum_{S\in H} X^S(0) = 0$$

Al sustraer 4.1 de la definición dada de $g_k$, obtenemos,

$$g_k = \sum_{i=1}^{k}(-1)^i h_{k-i}(x_k, \cdots, x_n)(y_i - \sigma_i) \tag{4}$$

lo cual prueba que $< g_1, \cdots, g_n > \subset < \sigma_1 - y_1, \cdots, \sigma_n - y_n >$.
Para mostrar la otra inclusión hay que notar que, ya que $h_0 = 1$, se puede escribir 4.2 como

$$g_k = (-1)^k(y_k - \sigma_k) + \sum_{i=1}^{k-1}(-1)^i h_{k-i}(x_k, \cdots, x_n)(y_i - \sigma_i) \tag{5}$$

Para mostrar $\langle \sigma_1 - y_1, \cdots, \sigma_n - y_n \rangle \subset \langle g_1, \cdots, g_n \rangle$ basta observar que de (4.3) se tiene que

$$\sigma_k - y_k = (-1)^{1-k} g_k + (-1)^{2-k}\textstyle\sum_{i=1}^{k-1}(-1)^i h_{k-i}(x_k, \cdots, x_n)(y_i - \sigma_i)$$

y por inducción sobre $k$ se muestra que
si $k = 1$, $\quad \sigma_1 - y_1 = g_1$
si $k = 2$, $\quad \sigma_2 - y_2 = -g_2 + h_1(x_2, \cdots, x_n)g_1$
supongamos que para $k = s$,

$$\sigma_s - y_s = g_s - h'_{s-1}(x_s, \cdots, x_n)g_{s-1} + \cdots + h'_1((x_s, \cdots, x_n)g_1.$$

Luego,

$$\begin{aligned}\sigma_{s+1} - y_{s+1} &= (-1)^{-s} g_{s+1} + (-1)^{1-s} \sum_{i=1}^{s} (-1)^i h_{s+1-i}(x_{s+1}, \cdots, x_n)(y_i - \sigma_i) \\ &= (-1)^{-s} g_{s+1} + (-1)^{1-s}[(-1) h_{s+1}(x_{s+1}, \cdots, x_n)(y_1 - \sigma_1) + \cdots + \\ &\quad (-1)^s h_1(x_{s+1}, \cdots, x_n)(y_s - \sigma_s)] \\ &= (-1)^{-s} g_{s+1} + (-1)^{1-s}[(-1) h_{s+1}(x_{s+1}, \cdots, x_n) g_1 + \cdots + \\ &\quad (-1)^s h_1(x_{s+1}, \cdots, x_n) g_s - h_{s-1}^{'}(x_s, \cdots, x_n) g_{s-1} + \cdots + \\ &\quad h_1^{'}(x_s, \cdots, x_n) g_1].\end{aligned}$$

Por tanto, si $f \in \langle \sigma_1 - y_1, \cdots, \sigma_n - y_n \rangle$ entonces $f \in \langle g_1, \cdots, g_n \rangle$.
Por último, veamos que $lt(g_k) = x_k^k$. En efecto

$$\begin{aligned}lt(g_k) &= lt[h_k(x_k, \cdots, x_n) + \sum_{i=1}^{k} (-1)^i h_{k-i}(x_k, \cdots, x_n) y_i], k = 1, \cdots, n \\ &= \text{máx}[lt(h_k(x_k, \cdots, x_n), lt[\sum_{i=1}^{k} (-1)^i h_{k-i}(x_k, \cdots, x_n) y_i]], k = 1, \cdots, n) \\ &= \text{máx}[lt(h_k(x_k, \cdots, x_n), \text{máx}[lt(-h_{k-1}(x_k, \cdots, x_n) y_1), lt(h_{k-2}(x_k, \cdots, x_n) y_2), \cdots, \\ &\quad lt((-1)^k h_0(x_k, \cdots, x_n)) y_k]] \\ &= \text{máx}[lt(h_k(x_k, \cdots, x_n)), \text{máx}[lt(\sum_{|\alpha|=k} x_k^{\alpha_k} x_{k+1}^{\alpha_{k+1}} \cdots x_n^{\alpha_n}), lt(\sum_{|\alpha|=k-1} x_k^{\alpha_k} x_{k+1}^{\alpha_{k+1}} \cdots x_n^{\alpha_n} y_1), \\ &\quad lt(\sum_{|\alpha|=k-2} x_k^{\alpha_k} x_{k+1}^{\alpha_{k+1}} \cdots x_n^{\alpha_n} y_2), \cdots, lt((-1)^k y_k)]] \quad \textit{donde} \quad \alpha = \alpha_k + \alpha_{k+1} + \cdots + \alpha_n \\ &= \text{máx}(x_k^k, \text{máx}(x_k^{k-1}, x_k^{k-2}, \cdots, 1)) \\ &= x_k^k\end{aligned}$$

Así, los términos principales de $g_1, \cdots, g_n$ son primos relativos. Por la proposición 2.9 se obtiene $S(g_i, g_j) \xrightarrow{G}_+ 0$ y usando el teorema 2.8 concluimos que $\{g_1, \cdots, g_n\}$ forman una base de Gröbner para $\langle \sigma_1 - y_1, \cdots, \sigma_n - y_n \rangle$ □

**Algoritmo para verificar si polinomio es simetrico**

**ENTRADA**: $f \neq 0 \in k[x_1, \cdots, x_n]$
$\sigma_1, \cdots, \sigma_n$ las funciones simétricas elementales.
**SALIDA**: VERDADERO, si el polinomio es simétrico, y en este caso,
$f = g(\sigma_1, \cdots, \sigma_n)$
FALSO en otro caso.
**INICIO**: Calcule una base de Gröbner $G$ para $\langle \sigma_1 - y_1, \cdots, \sigma_n - y_n \rangle$
Calcule el residuo $g$ de la división de $f$ por la base de Gröbner.

**SI** $g \in k[y_1, \cdots, y_n]$ **ENTONCES**

resultado:= VERDADERO

$f = g(\sigma_1, \cdots, \sigma_n)$

**EN CASO CONTRARIO**
resultado:= FALSO

**RETORNE** resultado

**Ejemplo 4.3.** *Considere el polinomio*

$$f = x^3 + y^3 + z^3 \in \mathbb{Q}[x, y, z]$$

*Verifique si $f$ es simétrico, si es así, escríbalo en términos de los polinomios simétricos elementales $\sigma_1, \sigma_2, \sigma_3$, usando el orden lex en $\mathbb{Q}[x, y, z, y_1, y_2, y_3]$ con $x > y > z > y_1 > y_2 > y_3$.*

*Paso 1.* Calcular una base de Gröbner para $< \sigma_1 - y_1, \sigma_2 - y_2, \sigma_3 - y_3 >$.

Para esto usamos la proposición 4.2 y obtenemos que

$$\begin{aligned} g_1 &= h_1(x,y,z) + \sum_{i=1}^{1}(-1)^i h_{1-i}(x,y,z)y_i = \sum_{|\alpha|=1} \mathbf{x}^\alpha + (-1)h_0(x,y,z)y_1 \\ &= x + y + z - y_1 \\ g_2 &= h_2(y,z) + \sum_{i=1}^{2}(-1)^i h_{2-i}(y,z)y_i = \sum_{|\alpha|=2} \mathbf{x}^\alpha + (-1)h_1(y,z)y_1 + h_0(y,z)y_2 \\ &= y^2 + yz - yy_1 + z^2 - zy_1 + y_2 \\ g_3 &= z^3 - z^2y_1 + zy_2 - y_3 \end{aligned}$$

forman una base de Gröbner para el ideal $\langle \sigma_1 - y_1, \sigma_2 - y_2, \sigma_3 - y_3 \rangle \subset \mathbb{Q}[x,y,z,y_1,y_2,y_3]$. Así,

$$G = \{x + y + z - y_1, y^2 + yz - yy_1 + z^2 - zy_1 + y_2, z^3 - z^2y_1 + zy_2 - y_3\}$$

*Paso 2.* Aplicar el algoritmo de la división para hallar el residuo $g$ obtenido al dividir $f$ entre $G$. Este residuo fue calculado en [6] usando el algoritmo de la división dado en [1]. Usando el programa CoCoa.

$UseR ::= \mathbb{Q}[x,y,z,y_1,y_2,y_3]$;
$F := x^3 + y^3 + z^3$;
$L := [x + y + z - y_1, y^2 + yz - yy_1 + z^2 - zy_1 + y_2, z^3 - z^2y_1 + zy_2 - y_3]$;
$DivAlg(F, [x + y + z - y_1, L])$;
$Record[Quotients = [x^2 - xy + y^2 - xz + 2yz + z^2 + xy_1 - 2yy_1 -$
$2zy_1 + y_1^2, -3z + 3y_1, 3], Remainder = y_1^3 - 3y_1y_2 + 3y_3]$

luego,

$$g = y_1^3 - 3y_1y_2 + 3y_3$$

*Paso 3.* $g(y_1,y_2,y_3) = y_1^3 - 3y_1y_2 + 3y_3 \in \mathbb{Q}[y_1,y_2,y_3]$, y por tanto $f$ es simétrico. Así,

$$f = g(\sigma_1,\sigma_2,\sigma_3) = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3.$$

## 5. APLICACIÓN DE POLINOMIOS SIMÉTRICOS EN LA TEORÍA DE INVARIANTES

El ejemplo más básico de invariantes de un grupo finito de matrices está dado por los polinomios simétricos, al considerar el grupo finito de las matrices de permutación, como ilustraremos a continuación.

**Definición 5.1.** *El grupo $S_n$ de todas las matrices cuadradas de tamaño $n$ cuyas entradas son 0 ó 1, pero de tal manera que hay un único 1 en cada fila y en cada columna se llama el grupo de matrices de permutación.*

**Ejemplo 5.2.** *Sea $S_3 \subset GL(3,k)$ el grupo de matrices de permutación.*

$$S_3 = \left\{ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \right\}$$

*Sea $f \in k[x,y,z]^{S_3}$ entonces $f(x) = f(A \cdot x), \forall A \in S_3$; ya que,*

$$f\left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right] = f(x,y,z) \qquad f\left[\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right] = f(y,z,x)$$

$$f\left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right] = f(x,z,y) \qquad f\left[\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right] = f(z,x,y)$$

$$f\left[\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right] = f(y,x,z) \qquad f\left[\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right] = f(z,y,x)$$

*se obtiene que, $f(x,y,z) = f(x,z,y) = f(y,x,z) = f(y,z,x) = f(z,x,y) = f(z,y,x)$. Los polinomios que cumplen esta condición son los polinomios simétricos, por tanto, los polinomios invariantes son los polinomios simétricos en $k[x,y,z]$.*

**Ejemplo 5.3.** *En general si se considera el grupo $S_n \subset GL(n,k)$ de matrices de permutación, entonces*

$$k[x_1, \cdots, x_n]^{S_n} = \{\textit{cualquier polinomio simétrico en} \quad k[x_1, \cdots, x_n]\} \tag{6}$$

*Por teorema (3.6), se conoce que los polinomios simétricos son polinomios en las funciones simétricas elementales con coeficientes en $k$, por tanto, se puede escribir (5.1) como,*

$$k[x_1, \cdots, x_n]^{S_n} = k[\sigma_1, \cdots, \sigma_n]$$

*Así, cualquier invariante puede ser escrito como un polinomio en las funciones simétricas elementales $\sigma_1, \cdots, \sigma_n$, además la representación en términos de las funciones simétricas elementales es única, por lo tanto, se obtiene un conocimiento explícito de los invariantes de $S_n$.*

## BIBLIOGRAFÍA

[1] Adams W. and Loustaunau P. (1994). *An Introduction to Gröbner Bases*. (Graduate Studies in Mathematics, Vol 3). Providence, R.I: American Mathematical Society.

[2] Cox D. and Little J. and O'Shea D. (1997). *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. New York: Springer.

[3] Hungerford T.W. (1974). *Algebra*. New York: Springer-Verlag.

[4] Kreuzer M. and Robbiano L. (2000). *Computational Commutative Algebra 1*. Berlin: Springer-Verlag.

[5] Kreuzer M. and Robbiano L. (2000). *Computational Commutative Algebra 2*. Berlin: Springer-Verlag.

[6] Patiño B. y Pérez H. (2009). *Uso de las bases de Gröbner en el cálculo de invariantes de grupos finitos*. Trabajo de grado. Tunja: Universidad Pedagógica y Tecnológica de Colombia.

[7] Sturmfels B. (2008). *Algorithms in Invariant Theory*. Berlin: Springer-Verlag. Second edition.



| Referencia | Fecha de recepción | Fecha de aprobación |
|---|---|---|
| Cifuentes, V., Patiño, B., Pérez, H.<br>Las bases de Gröbner en el estudio de los polinomios simétricos.<br>Revista *Tumbaga* (2010), 5, 195-210. | Día/mes/año<br>05/10/2009 | Día/mes/año<br>03/11/2009 |

# Evaluación de la efectividad del método de Murphy para la interpretación de señales en el gráfico de control multivariado $T^2$

Alex J. Zambrano[I], Hector F. López[II]

**Resumen.** Frecuentemente puede ser problemático identificar cuando una característica o grupo de características de calidad presenta un proceso fuera de control en el control multvariado de procesos estadísticos, ya que la calidad es representada por el control simultáneo de varias variables aleatorias correlacionadas.
En este trabajo se estudia la metodología de Murphy para detectar las causas asignables en una señal de fuera de control en el gráfico de control multivariado $T^2$. Se propone evaluar la efectividad de este método, utilizando simulación de procesos multivariados, cuando ocurren cambios en el vector de medias, e igualmente cuando las variables se encuentran correlacionadas. Se presenta un gráfico que muestra los resultados obtenidos.

**Palabras clave:** Control de calidad multivariado, estadístico $T^2$ de Hotelling, gráfico de control $T^2$, selección de variables, análisis discriminante.

**Abstract.** To identify when a characteristic or group of characteristics of quality presents a process out of control in the multivariate statistical process control can be a problem, because the quality is represented by the simultaneous control of correlated random variables.
In this paper we study the methodology of Murphy to detect assignable causes in an out of control signal in $T^2$ control chart. The purpose is to evaluate the effectiveness of this method, using multivariate process simulation, when occur changes in the mean vector, and when the variables are correlated. A graph shown the results obtained.

**Key words:** Multivariate quality control, Hotelling's $T^2$ statistic, $T^2$ control chart, selection of variables, discriminant analysis.

I Estudiante de Maestría en Estadística, Universidad Nacional de Colombia, Bogota, Colombia, ajzambranoc@unal.edu.co.

II Profesional en Matemáticas con énfasis en Estadística, Universidad del Tolima, Ibagué, Colombia, heticor86@hotmail.com.

## 1. INTRODUCCIÓN

Según Bersimis, Psarakis, y Panaretos (2007) los gráficos de control multivariados son una herramienta poderosa en el Control de Proceso Estadístico en cuanto a la identificación en un proceso fuera de control. Woodall y Montgomery (1999) enfatizan la necesidad de investigar más en esta área dado que la mayoría de procesos involucra un gran número de variables que son correlacionadas. Jackson (1991) comenta que cualquier procedimiento en control de calidad multivariado debe cumplir las siguientes condiciones:

1. Responder la pregunta "¿Esta el proceso en control?".
2. La probabilidad conjunta para el error tipo I debe especificarse.
3. Las relaciones entre las variables deben tenerse en cuenta.
4. Se debe dar un procedimiento que permita responder la pregunta "Si el proceso está fuera de control, ¿Cuál es el problema?".

La última pregunta ha provisto un interesante tema para muchos investigadores en los últimos años. Woodall y Montgomery (1999) manifiestan que aunque hay dificultad en la interpretación de señales en los gráficos de control multivariado se ha trabajado en métodos de reducción y técnicas gráficas. Maravelakis (2003) describe algunos de estos trabajos, entre ellos destacamos los realizados por Alt (1985) quien trabaja usando $p$ gráficos de control univariado con límites Bonferroni; Jackson (1991) propone usar la región de control elíptica y también componentes principales, y Mason, Tracy, y Young (1995, 1997) usan la descomposición del estadístico $T^2$ en partes independientes (llamada descomposición MYT) las cuales reflejan la contribución de las variables de manera individual y condicional. De este último trabajo algunas propuestas por otros autores incluyen esta descomposición, entre ellos Roy (1958), Murphy (1987), Doganaksoy, Faltin, y Tucker (1991), Hawkins (1991, 1993), Timm (1996) y Runger, Alt, y Montgomery (1996). La metodología de Murphy (1987) consiste en utilizar el Análisis Discriminante con el cual se detecta la variable o el conjunto de variables que hacen que el proceso presente una señal fuera de control. En trabajos anteriores se evaluó la efectividad de la descomposición MYT Zambrano y Zambrano (2008), para este se pretende evaluar la metodología de Murphy (1987) para la interpretación de señales.

Este artículo pretende abordar los conceptos básicos del gráfico de control multivariado $T^2$ para una observación individual (sección 2), consecuentemente se

describe la metodología propuesta por Murphy (1987) (sección 3), el procedimiento para evaluar esta metodología (sección 4), los resultados obtenidos (sección 5) y las conclusiones de los mismos (sección 6).

## 2. GRÁFICO DE CONTROL $T^2$

Este gráfico propuesto por Hotelling (1947) permite monitorear la distancia estadística entre un vector de observaciones y el vector de medias, para los casos en que se utilizan observaciones individuales. Suponiendo que los datos se distribuyen normal $p$ variada con vector de medias $\boldsymbol{\mu}_0$ y matriz de covarianzas $\boldsymbol{\Sigma}$ conocidos, la estadística $T^2$ para la $i$-ésima observación individual $\boldsymbol{X}_i^t = (X_{i1}, \ldots, X_{ip})$ viene dada por

$$T^2(\boldsymbol{X}_i) = (\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \, \boldsymbol{\Sigma}^{-1} \, (\boldsymbol{X}_i - \boldsymbol{\mu}_0) \quad i = 1, 2, \ldots, n. \tag{1}$$

Si se desconocen los parámetros de la normal $p$ variada, se utilizan sus estimaciones $\overline{\boldsymbol{X}}$ y $\boldsymbol{S}$ obtenidas a partir de datos históricos, por lo cual la ecuación (1) queda expresada como

$$T^2(\boldsymbol{X}_i) = \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^t \boldsymbol{S}^{-1} \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right) \quad i = 1, 2, \ldots, n, \tag{2}$$

donde el vector de medias muestrales viene dada por

$$\overline{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i,$$

y la matriz de covarianzas muestrales

$$\boldsymbol{S} = \frac{1}{n-1} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^t.$$

El gráfico de control consiste en gráficar la estadística $T^2$ v.s. el número de observaciones. Sin embargo, dependiendo de si se conocen o no los parámetros de la distribución normal $p$ variada, se considera una distribución diferente a un nivel de significancia $\alpha$ Díaz (2002). Está distribución, según el caso, permite asociar un límite de control superior el cual se representa en el gráfico de control como una línea recta con la cual se puede indicar si un proceso está bajo control. Está es una diferencia notable frente a los gráficos de control univariados ya que no tienen importancia el límite central ni el de control inferior debido a que la estadística $T^2$ por ser cuadrática nunca es negativa. Además, no presenta ningún

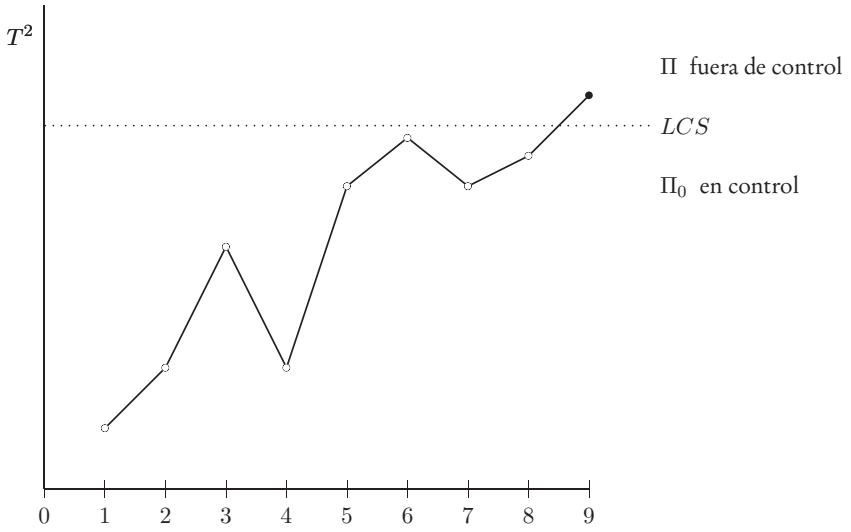



**Figura 1:** Gráfico de control $T^2$ con límite de control superior $LCS$.

problema si es muy cercano a cero (indicaría que la observación es muy cercana al vector de medias de las variables) (Figura 1).

Si la variabilidad de las observaciones en el proceso es natural o común, se presenta un gráfico de control donde cada estadística está por debajo del límite de control superior, se dice entonces que el proceso esta bajo control. En caso contrario se dice que el proceso esta fuera de control, y se deberá a que la variabilidad de las observaciones en el proceso no es natural o presenta alguna causa asignable o especial.

El gráfico de control $T^2$ se aplica en dos fases que se conocen como fase I y II; en la primera de ellas se toma un conjunto de datos históricos y se construye el gráfico de control $T^2$ para establecer si el proceso se encontraba bajo control cuando se sacaron $n$ observaciones individuales preliminares y se calcularon los estadísticos $\overline{\boldsymbol{X}}$ y $\boldsymbol{S}$. El objetivo en la fase I es obtener un conjunto de observaciones homogéneas bajo control a fin de establecer el límite de control superior para la fase II, el cual consiste en el monitoreo de la producción futura. En ocasiones se le denomina análisis retrospectivo Montgomery (2004).

### 2.1 Límites de control para la fase I

La fase I consta de dos etapas. La etapa 1 consiste en construir a partir de un conjunto de datos histórico el límite de control superior, para determinar si el proceso está bajo control. En esta etapa la distribución de la estadística $T^2$ dada por la ecuación (2) se distribuye

$$T^2 \sim \frac{(n-1)^2}{n} B_{(\frac{p}{2}, \frac{n-p-1}{2}, \alpha)}.$$

Para determinar si el proceso se encuentra bajo control se establece el límite de control superior asociado a la distribución del estadística anterior dada por

$$LCS = \frac{(n-1)^2}{n} B_{(\frac{p}{2}, \frac{n-p-1}{2}, \alpha)},$$

donde $B_{(\delta_1, \delta_2, \alpha)}$ es el percentil $(1-\alpha)$ de la distribución beta con parámetros $\delta_1$ y $\delta_2$. Cuando una observación hace que el proceso se encuentre fuera de control a está se la conoce como señal. Si el gráfico $T^2$ indica que hay una señal, se debe llevar a cabo una investigación con el objeto de encontrar causas especiales que la hubieran producido. Las observaciones multivariadas que correspondan a causas especiales se eliminan. Entonces se calculan de nuevo el límite de control superior para un examen retrospectivo basado en las observaciones restantes y el procedimiento se repite.

En la etapa 2 se prueba si el proceso permanece bajo control cuando nuevas observaciones sean seleccionadas. La estadística $T^2$ sigue una distribución $F$, por lo cual el gráfico de control $T^2$ utiliza en esta etapa un límite de control superior dado por,

$$LCS = \frac{p(n+1)(n-1)}{n(n-p)} F_{(p, n-p, \alpha)},$$

donde $n$ es el tamaño del observaciones en el conjunto de datos históricos con las que finalmente se estimaron los parámetros en la etapa 1 y $F_{(p, m-p, \alpha)}$ es el percentil $(1-\alpha)$ de la distribución $F$ con parámetros $p$ y $n-p$. Una vez estabilizado el proceso en la etapa 2, se asume que los estimadores finales $\overline{\boldsymbol{X}}$ y $\boldsymbol{S}$ son los verdaderos valores de los parámetros bajo control Vargas (2006).

### 2.2 Límites de control para la fase II

Para la fase II la carta $T^2$, dado que se conocen los parámetros la estadística $T^2$ se construyen como la ecuación (1), en este punto el límite de control superior

viene dado por

$$LCS = \chi^2_{(p,\alpha)}, \tag{3}$$

donde $\chi^2_{(p,\alpha)}$ es el percentil $(1-\alpha)$ de la distribución $\chi^2$ central con parámetro $p$.

Un desarrollo de los límites cuando se consideran subgrupos de observaciones mayores que uno se realiza de manera similar en Vargas (2006). En adelante se asumirá que el proceso ha pasado la Fase I, por lo que el proceso se encuentra bajo control.

## 3. METODOLOGÍA DE MURPHY

Este procedimiento es propuesto por Murphy (1987), subcaso de la descomposición de MYT según Maravelakis (2003). Se asume que el proceso está en fase II descrito anteriormente (ver sección 2), por lo cual cada observación individual $\boldsymbol{X}_i$ se distribuye normal $p$ variada con vector de medias $\boldsymbol{\mu}_0$ y matriz de covarianza $\boldsymbol{\Sigma}$. Estas ideas desarrolladas son extendibles para el caso donde $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ puedan ser estimados Murphy (1987).

La población bajo control se denotará por $\Pi_0 \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$. La familia de poblaciones fuera de control se denotara por $\Pi \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. El procedimiento hace monitoreos cuando hay algún cambio en la media $\boldsymbol{\mu}_0$.

Para tomar la decisión sobre i el proceso está o no bajo control, se utiliza la estadístico $T^2$ para una observación individual $\boldsymbol{X}_i$ dado por la ecuación (1) la cuál es equivalente a verificar la hipótesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ v.s. $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. Seleccionando el límite de control superior ($LCS$) correspondiente a la ecuación (3), se define la siguiente regla de decisión

Si $T^2 \leq LCS$; entonces $\boldsymbol{X}_i \in \Pi_0$ (proceso bajo control).

Si $T^2 > LCS$; entonces $\boldsymbol{X}_i \notin \Pi_0$ (proceso fuera de control).

### 3.1 Ventajas de utilizar el gráfico de control $T^2$

La ventaja principal de utilizar estadístico $T^2$ en el gráfico de control $T^2$, es la propia reflexión sobre la estructura de correlación de las poblaciones (están implícitas las correlaciones de las poblaciones, ya que $\boldsymbol{\Sigma}$ es conocida), aunque también se presenta la misma situación con un buen estimador de $\boldsymbol{\Sigma}$. En la práctica, el no tener en cuenta la correlación entre las variables es una desventaja puesto se deben utilizar cartas independientes para monitorear cada una de las $p$ características de calidad de manera individual. Haciendo esto se hacen pruebas

simultáneas sobre cartas individuales permitiendo que cada error de tipo I sea $\alpha/p$, en lugar de $\alpha$, lo cual es una dificultad.
Otra razón para utilizar el gráfico de control $T^2$ es la facilidad de cálculo y de construcción simple del mismo ya que sólo requiere comparar valores $T^2$ con el límite de control superior (Figura 1).

### 3.2 Interpretación de una señal y selección de la variable causante

Se propone utilizar el Análisis Discriminante Díaz (2002). Se plantea interpretar una señal fuera de control y una prueba sobre cuál de las variables (o subconjunto de ellas) causaron la señal.
Dada una observación $\boldsymbol{X_*}$ que produce una señal en el gráfico de control $T^2$ ($T^2(\boldsymbol{X}_*) > LCS$), la pregunta de interés inmediato es: ¿Cuál de las $p$ variables, o subconjunto $p_1$ de ellas ($p = p_1 + p_2$) causaron la señal? Murphy (1987) señala que particionando la observación $\boldsymbol{X}_*^t = [\boldsymbol{X}_*^{(1)}, \boldsymbol{X}_*^{(2)}]$, donde $\boldsymbol{X}_*^{(1)}$ es el subconjunto de las $p_1$ variables las cuales se sospecha causaron la señal, y $\boldsymbol{X}_*^{(2)}$ es el conjunto de las $p_2$ variables restantes.
Además, denotando con $T_p^2$ la distancia cuadrada completa,

$$T_p^2 = T^2\left(\boldsymbol{X}_*\right) = \left(\boldsymbol{X}_* - \boldsymbol{\mu}_0\right)^t \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{X}_* - \boldsymbol{\mu}_0\right)$$

y con $T_{p_1}^2$ la distancia reducida correspondiente al subconjunto $p_1$,

$$T_{p_1}^2 = T^2\left(\boldsymbol{X}_*^{(1)}\right) = \left(\boldsymbol{X}_*^{(1)} - \boldsymbol{\mu_0}^{(1)}\right)^t \boldsymbol{\Sigma}_{11}^{-1}\left(\boldsymbol{X}_*^{(1)} - \boldsymbol{\mu_0}^{(1)}\right)$$

donde $\boldsymbol{\mu}_0$ y $\boldsymbol{\Sigma}$ están particionadas como lo esta $\boldsymbol{X}_*$. Finalmente, la siguiente diferencia es calculada

$$D = T_p^2 - T_{p_1}^2, \tag{4}$$

el cual $D \sim \chi_{p_1}^2$. Si $D$ es grande, se rechaza la hipótesis nula de que el subconjunto $p_1$ causó la señal. Si es pequeño, no se rechaza. Pruebas similares son bien conocidas en el análisis discriminante y en el análisis de regresión para tratar con problemas de selección de variables Díaz (2002). Si ($\boldsymbol{\mu_0}, \boldsymbol{\Sigma}$), son estimados, la prueba apropiada para $D$ es una distribución $F$ Seber (1984).

### 3.3 Algoritmo de Murphy para seleccionar variables fuera de control

El siguiente algoritmo describe y reduce el cálculo computacional para considerar señales fuera de control. Se consideran observaciones individuales. Si se desea trabajar con observaciones en subgrupos mayores que uno, este mismo algoritmo se describe en Murphy (1987).

**Paso 1** Llevar a cabo una prueba $T^2$ con un nivel de significancia $\alpha$. Si la condición se encuentra fuera de control entonces se continua con el paso 2.

**Paso 2** Calcular los $p$ individual $T_1^2(X_i)$ equivalente a mirar las $p$ cartas individuales y calcular las $p$ diferencias $D_{p-1}(i) = T_p^2 - T_1^2(X_i)$. Seleccionar el $\text{mín}\{D_{p-1}(i)\} = D_{p-1}(r)$ y probar esta mínima diferencia. Si $D_{p-1}(r)$ es no significante entonces la variable $r$-ésima es la única que requiere atención. Si $D_{p-1}(r)$ es significante continue con el paso 3.

**Paso 3** Calcular las $p-1$ diferencias $D_{p-2}(r,j) = T_p^2 - T_2^2(X_r, X_j)$, $1 \leq r, j \leq p$ y $r \neq j$. Seleccionar el $\text{mín}\{D_{p-2}(r,j)\} = D_{p-2}(r,s)$ y pruebe la mínima diferencia. Si $D_{p-2}(r,s)$ no es significante entonces las variables $r$-ésima y $s$-ésima son las únicas que requieren atención. Si $D_{p-2}(r,s)$ es significante entonces continue con el paso 4.

**Paso 4** Similar al paso 3.

**Paso .** Similar al paso 3 y 4.

**Paso p** Si la última $D_{p-(p-1)}$ es significante, entonces, todas las $p$ variables requieren atención.

Murphy (1987) recomienda que en las pruebas $D_{p-i} \sim \chi^2_{(p-i,\alpha^*)}$ se use un nivel de significancia en el intervalo $0.1 \leq \alpha^* \leq 0.2$.

**Ejemplo 3.1.** Considérese un proceso de tres variables, el cual se distribuye multivariadamente normal con parámetros poblacionales dados por el vector de medias ceros ($\boldsymbol{\mu} = \mathbf{0}$) y matriz de varianzas la idéntica de orden 3 ($\boldsymbol{\Sigma} = \boldsymbol{I}_3$), para observaciones individuales. Se calculará el estadístico $T^2$ para la observación $\boldsymbol{X}_i^t = [-0.1736, 2.9356, 0.0379]$.

$$T^2(\boldsymbol{X}_i) = (\boldsymbol{X}_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}) = 8.649026$$

Comparando este valor con el límite de control superior dado por la ecuación (3) para $\alpha = 0.05$ se observa que se presenta una señal ($LCS = \chi^2_{(3,0.05)} = 7.814728$).
Ahora según el algoritmo anterior calculan los términos individuales $T_1^2(X_i)$ ($i = 1, 2, 3$)

$$T_1^2(X_1) = 0.0314, \quad T_1^2(X_2) = 8.6177, \quad T_1^2(X_3) = 0.0014$$

Se calculan las diferencias utilizando la ecuación (4) y se elige la mínima diferencia de las tres anteriores ($D_2(2) = 0.0313$). Se determina que esta diferencia no

es significante ($\chi^2_{(2,0.1)} = 4.60517$), entonces la variable $X_2$ requiere atención ya que es la causante de la señal fuera de control.

## 4. METODOLOGÍA

Para evaluar la efectividad de la metodología de Murphy (1987) se propone estudiar cambios en las componentes de vector de medias y correlación entre variables tal como lo describimos a continuación.
Se propone generar observaciones individuales de un proceso multivariado normal estándar con máximo tres variables. Se toma una muestra de $n = 100$ bajo control. Como se conocen los parámetros entonces se considera el límite de control superior dado por (3) con un nivel de significancia de $\alpha = 0.05$. Se hacen cambios a las componentes del vector de medias, o a la matriz de covarianzas de la siguiente manera:

1. Se cambia una o dos o las tres componentes del vector de medias, aumentando está con valores $\mu = 0.5,\ 1,\ 1.5,\ 2,\ 3,\ 5$.

2. Se realizaran cambios a la matriz de covarianzas para obtener correlaciones entre las variables $X_1$ y $X_2$ del proceso con los siguientes valores $\rho_{12} = 0.1,\ 0.5,\ 0.8$. Con esto se espera observar sus efectos.

Luego de encontrar el proceso fuera de control en un tiempo $t$ se aplica la metodología de Murphy (1987) en este tiempo. Se interpretarán las causas que determinan la señal fuera de control. Esto se realizarán $5000$ veces para evaluar la efectividad de esta descomposición, en términos de probabilidades, lo que facilitará la interpretación de la misma.

## 5. RESULTADOS

La tabla 1 resume los resultados obtenidos de las diferentes corridas realizadas. La primera columna representa las correlaciones para las variables $(X_1, X_2)$, la segunda columna representa los valores de los corrimientos en vector de medias. Las siguientes columnas representan las probabilidades en las que el método de Murphy detecta adecuadamente las variables a las cuales se les realizó el cambio; e.d., si la columna 4 de la fila 4 hay un valor de $0.65$ indica que la probabilidad de que el método de Murphy detecte cuando se produce un corrimiento en la segunda componente del vector de medias de $\mu = 1.5$, cuando $\rho_{12} = 0$ es de un 65 %. A estos valores los llamaremos probabilidades de detención efectivas en las variables que producen señal. Recordemos que los corrimientos en

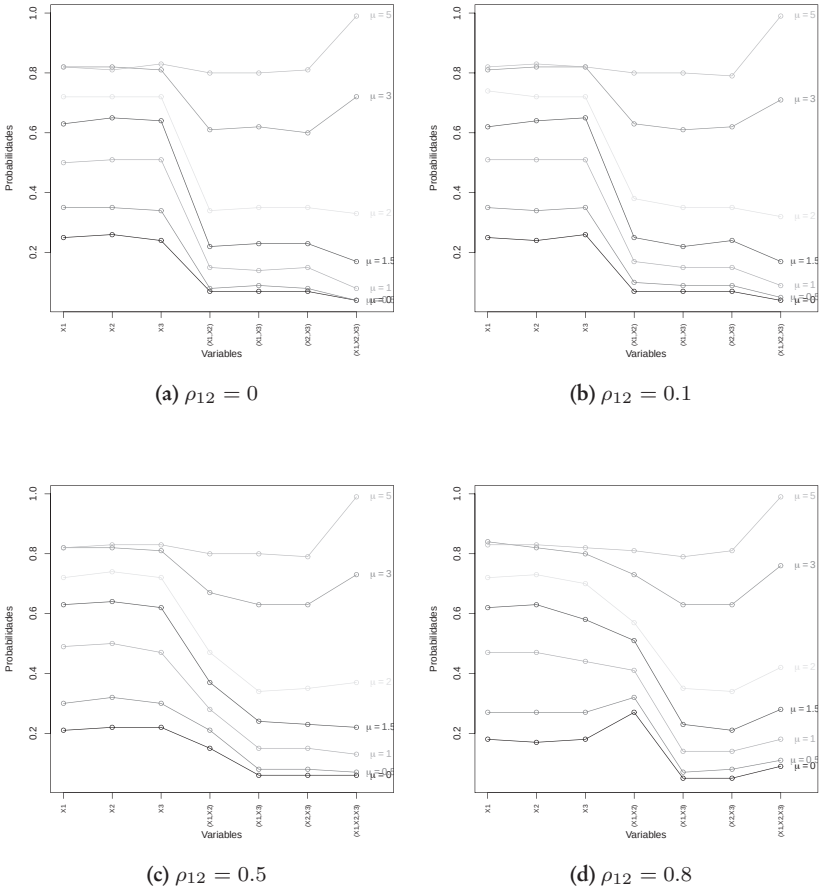
**Tabla 1:** Probabilidades de detección efectivas de las variables según corrimientos en el vector de medias y correlación entre la 1 y 2 variable

| $\rho_{12}$ | $\mu$ | $(\mu,0,0)$ | $(0,\mu,0)$ | $(0,0,\mu)$ | $(\mu,\mu,0)$ | $(\mu,0,\mu)$ | $(0,\mu,\mu)$ | $(\mu,\mu,\mu)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0.25 | 0.26 | 0.24 | 0.07 | 0.07 | 0.07 | 0.04 |
| | 0.5 | 0.35 | 0.35 | 0.34 | 0.08 | 0.09 | 0.08 | 0.04 |
| | 1 | 0.50 | 0.51 | 0.51 | 0.15 | 0.14 | 0.15 | 0.08 |
| | 1.5 | 0.63 | 0.65 | 0.64 | 0.22 | 0.23 | 0.23 | 0.17 |
| | 2 | 0.72 | 0.72 | 0.72 | 0.34 | 0.35 | 0.35 | 0.33 |
| | 3 | 0.82 | 0.82 | 0.81 | 0.61 | 0.62 | 0.60 | 0.72 |
| | 5 | 0.82 | 0.81 | 0.83 | 0.80 | 0.80 | 0.81 | 0.99 |
| 0.1 | 0 | 0.25 | 0.24 | 0.26 | 0.07 | 0.07 | 0.07 | 0.04 |
| | 0.5 | 0.35 | 0.34 | 0.35 | 0.10 | 0.09 | 0.09 | 0.05 |
| | 1 | 0.51 | 0.51 | 0.51 | 0.17 | 0.15 | 0.15 | 0.09 |
| | 1.5 | 0.62 | 0.64 | 0.65 | 0.25 | 0.22 | 0.24 | 0.17 |
| | 2 | 0.74 | 0.72 | 0.72 | 0.38 | 0.35 | 0.35 | 0.32 |
| | 3 | 0.81 | 0.82 | 0.82 | 0.63 | 0.61 | 0.62 | 0.71 |
| | 5 | 0.82 | 0.83 | 0.82 | 0.80 | 0.80 | 0.79 | 0.99 |
| 0.5 | 0 | 0.21 | 0.22 | 0.22 | 0.15 | 0.06 | 0.06 | 0.06 |
| | 0.5 | 0.30 | 0.32 | 0.30 | 0.21 | 0.08 | 0.08 | 0.07 |
| | 1 | 0.49 | 0.50 | 0.47 | 0.28 | 0.15 | 0.15 | 0.13 |
| | 1.5 | 0.63 | 0.64 | 0.62 | 0.37 | 0.24 | 0.23 | 0.22 |
| | 2 | 0.72 | 0.74 | 0.72 | 0.47 | 0.34 | 0.35 | 0.37 |
| | 3 | 0.82 | 0.82 | 0.81 | 0.67 | 0.63 | 0.63 | 0.73 |
| | 5 | 0.82 | 0.83 | 0.83 | 0.80 | 0.80 | 0.79 | 0.99 |
| 0.8 | 0 | 0.18 | 0.17 | 0.18 | 0.27 | 0.05 | 0.05 | 0.09 |
| | 0.5 | 0.27 | 0.27 | 0.27 | 0.32 | 0.07 | 0.08 | 0.11 |
| | 1 | 0.47 | 0.47 | 0.44 | 0.41 | 0.14 | 0.14 | 0.18 |
| | 1.5 | 0.62 | 0.63 | 0.58 | 0.51 | 0.23 | 0.21 | 0.28 |
| | 2 | 0.72 | 0.73 | 0.70 | 0.57 | 0.35 | 0.34 | 0.42 |
| | 3 | 0.84 | 0.82 | 0.80 | 0.73 | 0.63 | 0.63 | 0.76 |
| | 5 | 0.83 | 0.83 | 0.82 | 0.81 | 0.79 | 0.81 | 0.99 |

el vector de medias para que produzca una señal pueden ser corrimientos de manera individual $(\mu,0,0)$, $(0,\mu,0)$, $(0,0,\mu)$, corrimientos en dos componentes $(\mu,\mu,0)$, $(\mu,0,\mu)$, $(0,\mu,\mu)$ y corrimiento en las tres componentes $(\mu,\mu,\mu)$. Nótese que cuando $\rho_{12}=0$ y el corrimiento se hace en las tres componentes del vector de medias para $\mu=0$ la probabilidad de detectar una observación fuera de control cuando no se hicieron corrimientos es alrededor del 4 %, lo que se acerca bastante al nivel de significancia definido.

Realizando un gráfico de perfiles (Figura 2) se observa que las probabilidades de detección efectivas aumentan cuando el corrimiento en las medias es alto. Además se observa un comportamiento similar en los valores de probabilidad de detención en las corrimientos del vector de medias de manera individual,

independiente de la correlación $\rho_{12}$, por tanto no se afecta la efectividad de detención en las variables corridas de manera independiente. Así mismo cuando existe correlación baja o no la hay, se observan que los valores de probabilidad son similares (Figuras 2a y 2b). Por el contrario, cuando la correlación es moderada o grande, se observan un aumento en la probabilidad de detección cuando se realiza corrimientos simultáneos en las variables $(X_1, X_2)$ (Figuras 2c y 2d). Cuando se tiene una correlación alta se afectara la probabilidad de detección efectividad (Figura 2d).




**(a)** $\rho_{12} = 0$ **(b)** $\rho_{12} = 0.1$

**(c)** $\rho_{12} = 0.5$ **(d)** $\rho_{12} = 0.8$

**Figura 2:** Perfiles de las probabilidades efectivas de detección de las variables según corrimientos en el vector de medias y correlación

## 6. CONCLUSIONES

Al analizar las tablas se observan algunos comportamientos en los datos, que se ven reflejados a través de las probabilidades realizadas en la metodología de Murphy (1987). Estos tipos de comportamientos se analizarán para determinar las conclusiones:

1. Como se observó la metodología de Murphy (1987) es una buena alternativa para la detección de las causas de una señal fuera de control en un proceso multivariado.

2. La efectividad de detección de la variable que ocasiono la señal fuera de control es mayor cuando se ha realizado un cambio grande en el vector de medias. Al realizar cambios muy pequeños la probabilidad es baja.

3. Para correlaciones altas entre las variables $(X_1, X_2)$, la probabilidad de detección es alta.

4. La detención efectiva en las variables de manera independiente, cambia cuando el corrimiento en media es pequeño o no hay y la correlación es alta.

Para estudios posteriores se podría aumentar el numero de variables, y tamaños de subgrupos mayores que uno, para aplicar la metodología y determinar su efectividad ante estos cambios. El estudio de la metodología de Murphy (1987) se centra en utilizar análisis discriminante de Fisher, el cual compara las medias de dos poblaciones multivariadas distribuidas normal $p$-variadas con la misma matriz de covarianzas. Sin embargo, al utilizar nuestra metodología de estudio consideramos correlaciones entre variables, e.d., cambios en la matriz de covarianzas sin modificar la regla del discriminante de Fisher. Un posible estudio sería evaluar qué pasa si se cambia la regla del discriminante a matrices de covarianzas distintas Díaz (2002). Das y Prakash (2008) comparan las propuestas realizadas Mason y cols. (1995), Murphy (1987), Hawkins (1993) y Doganaksoy y cols. (1991) utilizando pruebas de potencia sin tener en cuenta la fase en la que se encuentra cada proceso. Se pueden observar metodologías diferentes utilizando pruebas empíricas como las realizadas en este trabajo, teniendo en cuenta la fase en la que se encuentra el proceso.

## BIBLIOGRAFÍA

Alt, F. B. (1985). Multivariate quality control. En S. Kotz, N. L. Johnson, y C. R. Read (Eds.), *The enciclopedia of statistical sciences* (pp. 110–122). New York: John Wiley & Sons.

Bersimis, S., Psarakis, S., y Panaretos, J. (2007). Multivariate statistical processs control charts: An overview. *Quality and Reliability Engineering International*, *23*(5), 517–543.

Das, N., y Prakash, V. (2008). Interpreting the out-of-control signal in multivariate control chart - a comparative study. *The International Journal of Advanced Manufacturing Technology*, *37*(9), 966–979.

Doganaksoy, N., Faltin, F. W., y Tucker, W. T. (1991). Identification of out of control multivariate characteristic in a multivariable manufacturing environment. *Communications in Statistics - Theory and Methods*, *20*(9), 2775–2790.

Díaz, L. G. (2002). *Estadística Multivariada: Inferencia y Métodos* (Primera ed.). Bogotá: Universidad Nacional de Colombia, Departamento de Matemáticas y Estadística.

Hawkins, D. M. (1991). Multivariate quality control based on regression adjusted variables. *Technometrics*, *33*(1), 61–75.

Hawkins, D. M. (1993). Regression adjustment for variables in multivariate quality control. *Journal of Quality Control*, *25*(3), 170–182.

Hotelling, H. (1947). Multivariate quality control - illustrated by the air testing of sample bombsights. En C. Eisenhart, M. W. Hastay, y W. A. Wallis (Eds.), *Techniques of statistical analysis* (pp. 111–184). New York: MacGraw Hill.

Jackson, J. E. (1991). *A user guide to principal components*. New York: Jhon Wiley & Sons.

Maravelakis, P. E. (2003). *An investigation of some characteristics of univariate and multivariate control charts*. Tesis Doctoral, Departament of Statistics, Athenas University of Economics and Business.

Mason, R. L., Tracy, N. D., y Young, J. C. (1995). Decomposition of $T^2$ for multivariate control chart interpretation. *Jornal of Quality Technology*,

*27*(2), 99–108.

Mason, R. L., Tracy, N. D., y Young, J. C. (1997). A practical approach for interpreting multivariate $T^2$ control chart signals. *Jornal of Quality Technology*, *29*(4), 396–406.

Montgomery, D. C. (2004). *Control Estadístico de la Calidad* (Tercera ed.). México: Limusa Wiley.

Murphy, B. J. (1987). Out of control variables with the $T^2$ multivariate quality control procedure. *Royal Statistical Society*, *36*, 571–581.

Roy, J. (1958). Step down procedure in multivariate analysis. *Annals of Mathematical Statistics*, *29*, 1177–1187.

Runger, G. C., Alt, F. B., y Montgomery, D. (1996). Contributors to a multivariate statistical process control chart signal. *Communications in Statistics. Theory and Methods*, *25*(10), 2203–2213.

Seber, G. A. F. (1984). *Multivariate observations*. New York: John Wyley.

Timm, N. H. (1996). Multivariate quality control using finite intersection tests. *Journal of Quality Control*, *28*(2), 233–243.

Vargas, J. A. (2006). *Intoducción al Control Estadístico de Calidad*. Bogotá: Universidad Nacional de Colombia, Departamento de Matemáticas y Estadística.

Woodall, W. H., y Montgomery, D. C. (1999). Research issues and ideas in statistical process control. *Journal of Quality Technology*, *31*(4), 376–386.

Zambrano, A. J., y Zambrano, L. A. (2008). Evaluando la efectividad de la descomposición MYT para la interpretación de señales fuera de control en la carta $T^2$. *Revista Tumbaga*, *3*, 141–157



| Referencia | Fecha de recepción | Fecha de aprobación |
|---|---|---|
| Zambrano, Alex J. y López, Héctor F.<br>Evaluando la efectividad de la descomposición MYT para la interpretación de señales fuera de control en la carta $T^2$.<br>Revista *Tumbaga* (2010), 5, pg 211-224 | Día/mes/año<br>16/07/2010 | Día/mes/año<br>02/09/2010 |

# Análisis de conglomerados en la identificación de estructura genética a partir de datos de marcadores moleculares

## Cluster analysis for identification of genetic structure from molecular marker data

Peña Malavera, Andrea;[I,II] Bruno, Cecilia;[I,III] Teich, Ingrid;[I,III]
Fernández, Elmer;[III,IV] Balzarini, Mónica.[I,III]

**Resumen**. En el contexto de abundante información genómica, como la producida a partir de marcadores moleculares basados en ADN, es de interés identificar la estructura genética subyacente en un conjunto de individuos, previo al análisis de asociación entre expresión de marcadores y fenotipo. Cuando existen subgrupos de individuos que difieren sistemáticamente en las frecuencias alélicas de sus marcadores, se origina una estructura genética que, de no ser considerada, incrementa el riesgo de detectar asociaciones espurias entre marcadores y fenotipo. Diversos métodos estadísticos son utilizados para determinar la agrupación de individuos desde datos de marcadores moleculares que producen información discreta multidimensional, entre ellos métodos basados en algoritmos de conglomerados jerárquicos (UPGMA), conglomerados no jerárquicos (K-means), redes neuronales como los mapas auto-organizativos (SOM) y métodos de conglomerados bayesianos. En este trabajo comparamos la capacidad de tales algoritmos para detectar subpoblaciones (conglomerados genéticos) bajo dos escenarios biológicos de estructura poblacional: modelo de islas y modelo de contacto. Los algoritmos de conglomerado fueron evaluados simultáneamente usando conjuntos de datos de marcadores moleculares de expresión binaria simulados bajo ambos modelos biológicos. El método de conglomeración bayesiano fue el que mejor identificó, entre los evaluados, las subpoblaciones simuladas bajo el modelo de migración de islas. Para el modelo de contacto la identificación de subgrupos fue difícil con cualquiera de los cuatro algoritmos de conglomeración evaluados.

**Palabras clave:** conglomerados jerárquicos, conglomerados no jerárquicos, mapas auto-organizativos, conglomerado bayesiano, modelos de migración.

**Abstract.** Prior to association studies, and in the context of abundant genomic information provided by molecular markers, it is of interest to identify the underlying genetic structure of individuals. Genetic structure arises when markers´ allele frequen-

I Estadística y Biometría-FCA, UNC
II FONCyT
III CONICET Córdoba- Argentina
IV Universidad Católica de Córdoba. Correo electrónico: andreapema@gmail.com

cies differ systematically between subgroups, and if it is not considered in association analysis, it increases the risk of detecting spurious associations between molecular markers and the phenotype of interest.

A variety of statistical methods are used to determine groups of individuals from molecular markers that produce multidimensional discrete data, such as methods based on hierarchical (UPGMA) and non-hierarchical clustering algorithms (K-means), neural networks (SOM), and Bayesian clustering. In this study, we compared the capacity of these algorithms to detect genetic clusters under two different biological scenarios: the island model and the contact model. The clustering algorithms were simultaneously evaluated using binary molecular marker data simulated under both biological scenarios. Bayesian clustering was the best model to identify subpopulations under the island migration model. However, in the contact model the identification of subgroups was difficult with all algorithms.

**Key words:** Hierarchical clusters, Non-hierarchical clusters, self-organizing maps, Bayesian clustering, Migration models.

## 1. INTRODUCCIÓN

En genética humana, como en genética de animales y plantas, es de interés identificar loci del ADN que rigen caracteres complejos (generalmente de expresión continua). El estudio de asociación se realiza en un contexto de abundante información genómica como es la producida por distintos tipos de marcadores moleculares del ADN usando individuos seleccionados al azar de una población con distinto nivel de relación de parentesco y estructuración genética. El procedimiento para el análisis simultáneo de información genómica y fenotípica con el propósito de identificar asociaciones no aleatorias entre loci de marcadores moleculares y loci del carácter fenotípico en estudio, es conocido como mapeo de asociación o mapeo por desequilibrio de ligamiento. Para este procedimiento se han utilizado con éxito distintos modelos estadísticos, los cuales deben contemplar si existen estructuras genéticas subyacentes en la población de estudio, es decir la posible existencia de subpoblaciones con distinto nivel de relación genómica. Diversos métodos analíticos computacionales de naturaleza exploratoria multivariada e índices de diversidad genética entre y dentro de grupos pueden usarse para detectar estructura genética (Hedrick, 2005).

Los estudios de asociación, suponen que si una mutación o cambio en el estado alélico del marcador incrementa la expresión de una característica en una población de individuos, entonces se puede esperar que el o los alelos asociados a tal característica sean más frecuentes entre los individuos que la comparten que entre el resto de los

individuos. Una población estructurada genéticamente es aquella en la que coexisten subgrupos o conglomerados de individuos que difieren sistemáticamente en sus frecuencias alélicas para los diferentes loci. Cuando se lleva a cabo un análisis de asociación en estas poblaciones sin considerar los efectos de la subestructura poblacional subyacente, se aumenta el riesgo de detectar asociaciones espurias entre marcadores y la característica fenotípica de interés.

El análisis estadístico de la información molecular a través de algoritmos de clasificación no supervisada, se utiliza para revelar tales estructuras de subpoblaciones que luego son usadas como un factor de clasificación dentro de los modelos de asociación. Diversos algoritmos estadístico-computacionales para la agrupación de entidades en espacios multidimensionales, tales como los métodos de conglomerados jerárquicos (UPGMA), conglomerados no jerárquicos (K-means), métodos de conglomeración bayesianos basados en cadenas de Markov Monte Carlo (MCMC), y redes neuronales que producen mapas auto-organizativos (SOM), pueden aplicarse a datos multialelos-multilocus para identificar subpoblaciones o conglomerados genéticos. El objetivo de este trabajo es ilustrar el desempeño relativo de estos métodos bajo distintos escenarios biológicos de presencia de subpoblaciones genéticas caracterizados por modelos de migración comunes en estudios de ecología y genética de poblaciones.

## 2. MATERIALES Y MÉTODOS

### 2.1 Datos

Se simularon datos genéticos multivariados del tipo binario con una estructura poblacional conocida a priori bajo dos escenarios biológicos caracterizados por diferentes modelos de migración: (A) modelo de islas y (B) modelo de contacto. El programa utilizado fue EASYPOP versión 2.0.1 (Balloux, 2001). Para el modelo de islas se consideró una estructura poblacional compuesta por 5 grupos, mientras que para el modelo de contacto se simuló una estructura poblacional de 2 grupos con 5 subpoblaciones cada uno (figura 1). El modelo de islas es considerado básico porque resume la estructura poblacional fundamental (Wright, 1931), y es utilizado en muchos estudios de genética de poblaciones para predecir procesos evolutivos causados por deriva génica, mutación, selección y migración (Chiappero et al., 2010). Este modelo asume que los migrantes tienen una frecuencia génica igual a la del conjunto de las subpoblaciones, y es considerado matemáticamente simple. El modelo de contacto se considera opuesto al de islas en el continuo de estructuras poblacionales. En él la

población no está dividida en subunidades donantes o receptoras de migrantes, ni es una unidad panmíctica. Los cruzamientos al azar están limitados por la distancia, de modo que los individuos tendrán una mayor probabilidad de aparearse con vecinos que con individuos más lejanos. De este modo se pueden agrupar a los individuos en «vecindarios», áreas definidas por «individuos centrales» cuyos progenitores se pueden tratar como extraídos al azar. La representación más sencilla del modelo de contacto es un hábitat lineal a lo largo del cual existe una distribución normal de las distancias entre los lugares de nacimiento de los padres y la progenie.

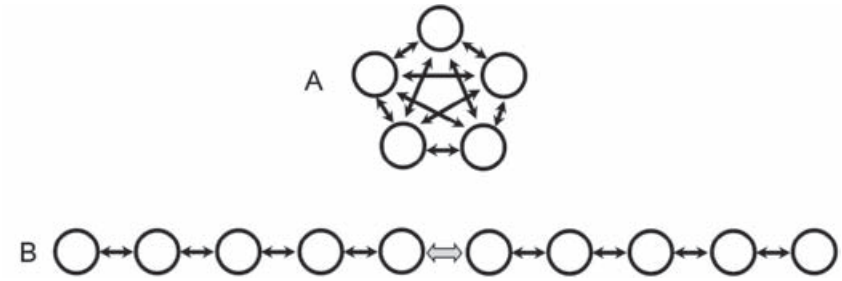



**Figura 1.** Representación esquemática de los dos modelos biológicos de migración simulados. A) Modelo de islas. B) Modelo de zona de contacto.

Los parámetros de simulación fueron fijados en este estudio con la finalidad de lograr datos moleculares que caractericen un conjunto de líneas o individuos genéticamente estructurados, y obtener escenarios comparables a la mayor parte de los estudios que se realizan. Se definieron los siguientes parámetros: nivel de ploidía (diploide), proporción de recombinación (0.01), número de individuos por población (100), número de loci (150), tasa de mutación (0.01), modelo de mutación (modelo de mutación paso a paso -SSM-), número de posibles estados alélicos (2), variabilidad de la población inicial (mínima), tasa de migración dentro de poblaciones (0.01), tasa de migración entre poblaciones (0.001). Todas las simulaciones fueron realizadas para 4000 generaciones.

El hecho de considerar una tasa de migración pequeña entre y dentro de poblaciones, significa que en cada generación existirá una baja proporción de la población que será sustituida por inmigrantes, tanto dentro de una misma población como entre poblaciones. Se espera que este modelo de migración provoque una alta endogamia local y elevada homocigosis entre individuos de una misma población, y por tanto mayor diferencia entre poblaciones, es decir que las poblaciones estén estructuradas genéticamente. Cada escenario o realización del modelo genético para los parámetros

expuestos anteriormente se simuló 100 veces y de cada conjunto de datos simulado se seleccionaron por muestreo aleatorio simple para retener en el análisis entre 120 y 180 individuos con igual proporción de representación de cada una de las subpoblaciones simuladas, en cada muestra emulada. El valor de la cantidad de líneas o individuos seleccionados fue elegido en función de los tamaños de muestras que se utilizan en la práctica en estudios de mapeo de asociación.

**2.2 Procedimientos evaluados**

Sobre cada conjunto de datos simulado se aplicaron simultáneamente 4 algoritmos de conglomerados: 1. Método jerárquico (UPGMA), 2. Método no jerárquico K-means, 3. Método basado en redes neuronales (SOM) y 4. Método bayesiano basado en MCMC.

En el método de conglomerados jerárquico UPGMA (Sokal y Michener, 1958), como en otros métodos de conglomerados de esta naturaleza, se parte de una matriz de distancias conteniendo todas las distancias entre pares de objetos, y se los comienza a agrupar teniendo en cuenta la mínima distancia. Luego de agrupar el primer par de individuos el proceso continua recalculando la matriz de distancias de manera tal que el conglomerado recientemente formado se trata como un nuevo objeto. Por esto es necesario definir una métrica de distancia entre objetos simples y conglomerados o entre conglomerados. Ésta se obtiene promediando todas las distancias entre pares de objetos, donde un miembro del par pertenece a uno de los conglomerados y el otro miembro al segundo conglomerado. Este es un método simple que se ha encontrado exitoso en numerosas aplicaciones de clasificación. La expresión para calcular la distancia promedio entre conglomerados es:

donde es la distancia entre el objeto, que pertenece al conglomerado AB y el objetoque pertenece al conglomerado C, siendo la sumatoria sobre todos los posibles pares de objetos entre dos conglomerados y donde y son los números de objetos en los conglomerados AB y C respectivamente. El método tiende a producir grupos de igual varianza (Milligan, 1980). La historia de formación de conglomerados en los distintos pasos del algoritmo de agrupamiento suele representarse gráficamente a través de un dendrograma.

El método de conglomerados no-jerárquico (K-means) agrupa objetos en k grupos haciendo máxima la variación entre conglomerados y minimizando la variación dentro de cada conglomerado. Este método comienza con un agrupamiento inicial o

con un grupo de puntos semilla (centroides) que formarán los centros de los grupos (partición inicial del grupo de objetos en k ítems), prosigue asignando cada objeto al grupo que tiene el centroide (media) más cercano. La partición lograda es aquella tal que la suma de la suma de las distancias al cuadrado de los miembros del grupo respecto a su centroide es mínima. El método se basa así en el principio de los k mejores centroides, éstos son modificados cada vez que un objeto se transfiere de un grupo al otro. El algoritmo K-means es óptimo en cada paso. Los resultados finales dependen de la configuración inicial, de la secuencia en que son considerados los objetos a agrupar y claramente del número de grupos. A los fines de alcanzar un óptimo global, es recomendable usar varias particiones iniciales y seleccionar aquella partición final con mínimo valor de la función objetivo.

Si bien estos métodos basados en conglomerados jerárquicos y no-jerárquicos se encuentran implementados en numerosos softwares, en este trabajo usamos la implementación del software Info-Gen (Balzarini y Di Rienzo, 2004). La función objetivo que se minimiza en Info-Gen es la suma de distancias al cuadrado entre los objetos de un mismo conglomerado. El gráfico que muestra la evolución de esta función objetivo para 2, 3,…, k conglomerados es usado para identificar el número de grupos subyacentes en la población de objetos, i.e. cantidad de grupos donde la función decrece a una menor tasa.

Los Mapas Auto-Organizativos o Self-Organizing Maps (SOM) son un modelo de red neuronal desarrollado por Kohonen (Kohonen, 1997). Este procedimiento procesa una base de datos o casos multidimensionales, resultando en un mapa (usualmente bidimensional) donde casos similares se "mapean" en regiones cercanas de la red neuronal. De esta manera "vecindad" significa "similaridad" (Fernández y Balzarini, 2007). La red se estructura como una capa usualmente bidimensional de nodos no conectados entre sí. Todos los nodos se asocian con un dato de entrada, se inicializan los pesos de cada nodo, se busca el nodo ganador respecto a su similitud con el caso de entrada, y se actualizan los pesos del nodo ganador y de sus vecinos, reiterando los pasos hasta que se satisface un criterio de detección impuesto previamente. Los mapas auto-organizativos constituyen un método de conglomeración similar a los métodos no jerárquicos donde los grupos que se conforman son ubicados espacialmente sobre la estructura de una red predefinida. Si bien originalmente fue implementado en Sompack (Kohonen, 1997) actualmente se puede implementar en los softwares Matlab, SAS y R, con distintos métodos de representación gráfica de los resultados. En el presente trabajo se usa el método de representación "U-matrix" disponible en Sompack.

El método de conglomeración bayesiano, basado en Cadenas de Markov de Monte Carlo propuesto por Pritchard y sus compañeros (2000) es frecuentemente usado para estudios de estructura genética poblacional. Se estima la distribución a posteriori de coeficientes asociados a cada individuo que se corresponden a los distintos subgrupos en los que éste puede clasificarse. El valor esperado de la distribución a posteriori provee una estimación de la proporción que el genoma de un individuo tiene o comparte con los distintos subgrupos. El algoritmo se encuentra implementado en el software STRUCTURE 2.2.3 (Pritchard et al., 2000).

### 2.3 Evaluación del desempeño de los procedimientos de conglomerados

Bajo cada escenario de modelo biológico de migración usado para simular estructura genética se contabilizó el porcentaje de corridas donde el algoritmo de conglomerado usado detectó correctamente el número de grupos o subpoblaciones simuladas (K). Adicionalmente se estimó para el conglomerado jerárquico UPGMA, no jerárquico K-means y el método bayesiano, el porcentaje de entidades bien clasificadas a través de la comparación de la subpoblación a la cual pertenece cada entidad (según simulación) y el conglomerado asignado. Para los mapas SOM se contabilizó el porcentaje de nodos en zonas de la red no claramente definidas como un conglomerado (transición). Estos porcentajes fueron calculados en aquellos escenarios donde se detectó con éxito el número de grupos de subpoblaciones (5 y 2 para el modelo de islas y de contacto, respectivamente). En las figuras 2 y 3 presentamos los gráficos obtenidos con cada técnica de conglomerados para un caso simulado bajo cada modelo biológico.

## 3. RESULTADOS

### 3.1 Desempeño de los algoritmos de agrupamiento a partir del modelo de Islas

En la figura 2 se representan los resultados obtenidos a partir de los 4 métodos de conglomerados. Cuando se utilizó el algoritmo de conglomerados jerárquico UPGMA, el dendrograma resultante permitió de manera relativamente fácil visualizar el agrupamiento de las líneas en 5 subgrupos con un umbral de corte para el eje de las abscisas (distancias entre conglomerados) posicionado aproximadamente en el percentil 95 de las distancias calculadas. En la gráfica obtenida luego de implementar un análisis de conglomerados no jerárquicos (K-means), se observó que la tasa de disminución del criterio que define la función objetivo (suma de cuadrados o heterogeneidad dentro de grupos) decae a partir de 5 conglomerados. La representación

de la red del mapa auto-organizativo debería permitir la visualización de 5 zonas con igual nivel de grises; sin embargo, en este ejemplo la observación de las mismas resultó difícil. Los puntos blancos representan nodos de la red que no han podido ser agrupados claramente en un conglomerado. En el método bayesiano se logró un gráfico en donde los 5 grupos fueron bien separados, sin embargo, en la clasificación de dicho algoritmo se observó la existencia de algunos individuos que perteneciendo a un grupo en particular, contienen características similares a las de otros conglomerados, lo que ocasiona en el gráfico líneas de individuos con más de un color.

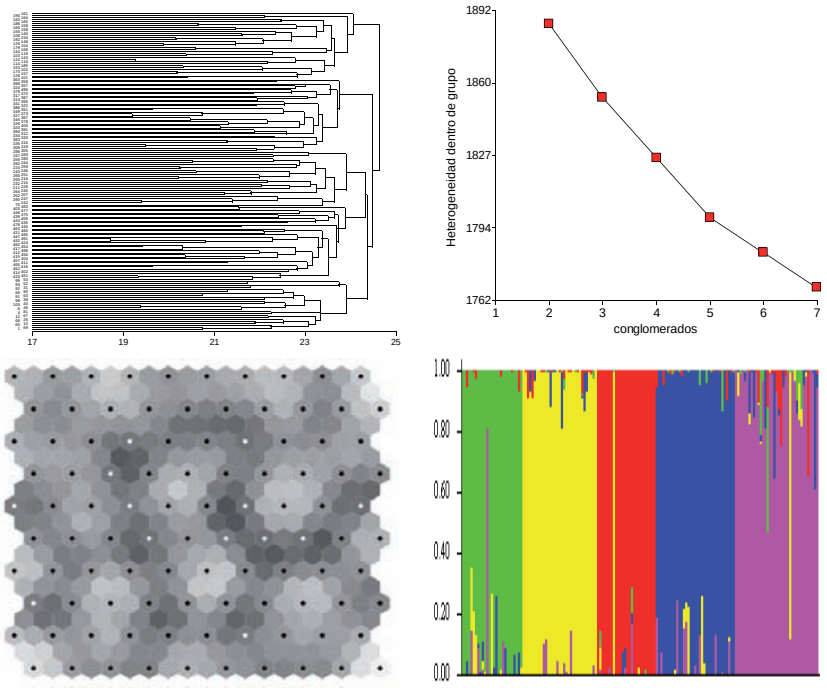



**Figura 2.** Representaciones gráficas de los resultados logrados por los algoritmos de agrupamiento (de izquierda a derecha, arriba) método jerárquico UPGMA, método no-jerárquico K-means, (abajo) redes neuronales basados en mapas auto-organizativos (SOM) y método bayesiano basados en MCMC, respectivamente. Modelo de islas con una tasa de migración dentro de poblaciones =0.01 y K=5.

### 3.2 Desempeño de los algoritmos de agrupamiento a partir del modelo de contacto

Las representaciones del modelo de contacto se presentan en la figura 3. La detección del número de grupos (K=2) fue relativamente fácil a través de los dendrogramas, al

igual que en el modelo de islas, mientras que la detección del agrupamiento de las líneas en 10 subgrupos (subpoblaciones) resultó difícil. En la representación gráfica obtenida a partir del análisis de conglomerados no jerárquicos se observa que la tasa de disminución del criterio que define la función objetivo no decae sustancialmente. La representación de la red del mapa auto-organizativo debería permitir la visualización de 10 zonas; en este ejemplo la observación de las mismas resultó difícil. Con el método bayesiano el gráfico resultante no permitió la identificación de los subgrupos (subpoblaciones) que conforman la estructura poblacional.

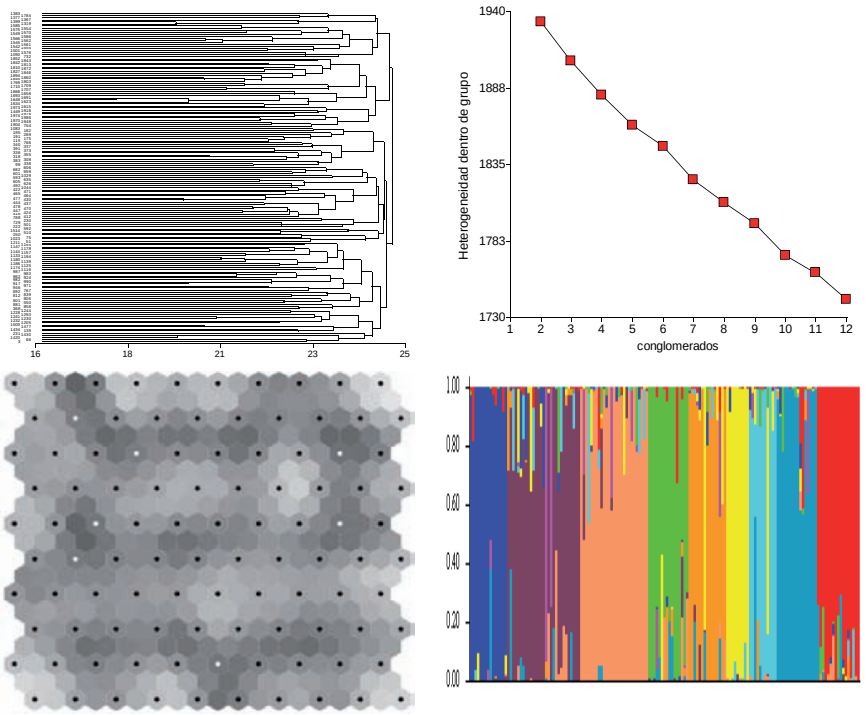



**Figura 3.** Representación gráfica de los resultados logrados por 4 algoritmos de agrupamiento (de izquierda a derecha, arriba) método jerárquico UPGMA, métodos no-jerárquico K-means, (abajo) redes neuronales basados en mapas auto-organizativos (SOM) y métodos bayesianos basados en MCMC, respectivamente. Modelo de contacto con una tasa migración dentro=0.01, tasa migración entre=0.001 y 2 grupos con 5 subpobaciones; total 10 subpoblaciones.

En la tabla 1 se presenta, para cada método de agrupamiento y para cada modelo de migración, la cantidad de veces que cada uno clasificó de manera correcta a cada individuo en el grupo o subpoblación que le correspondía según fuera simulado. Puede observarse que el método bayesiano fue el que presentó el mayor porcentaje

en la detección del número de grupo simulados (90%), seguido del método no jerárquico K-means que clasificó correctamente cada grupo el 83% de las veces para el modelo de islas. El método UPGMA y el SOM obtuvieron el número correcto de conglomerados el 56 y 23% de las veces, respectivamente. Esto último sugiere que tanto el método jerárquico como el basado en redes (SOM) son sensibles a modelos biológicos con estructuras más complejas. Con respecto a la clasificación correcta de individuos en los grupos bien identificados, los resultados estuvieron por encima del 90% para el modelo de islas en todos los métodos, excepto para el método SOM, que clasificó correctamente el 80% de los nodos. Con el método bayesiano el 99% de los individuos fue correctamente asignado en los grupos identificados. Los métodos UPGMA y K-means asignaron individuos correctamente en un 93% de las veces. Bajo el modelo de contacto, los 4 algoritmos evaluados presentaron una baja proporción de éxito en la detección del número de grupos (menor al 50%). El mayor porcentaje de éxito fue obtenido, a diferencia del modelo de islas, con el algoritmo de conglomerados jerárquicos, UPGMA.

Luego, a partir de los porcentajes de éxito en la detección del número de grupos o poblaciones (K), se contabilizó la proporción de individuos bien clasificados para cada grupo. Para los algoritmos jerárquico, bayesiano y SOM los porcentajes de entidades bien clasificadas fueron mayores al 80%. El método de conglomerados no jerárquico K-means tuvo en esta simulación bajo porcentaje de clasificación correcta de individuos (10%) cuando la estructura de población se obtuvo bajo un modelo biológico de contacto.

| Migración | Éxito en la detección | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | K-poblaciones+ | | | | Individuos @ | | | Nodos* |
| | UPGMA | Kmeans | SOM | Bayesiano | UPGMA | Kmeans | Bayesiano | SOM |
| Islas | 56 | 83 | 23 | 90 | 93 | 93 | 99 | 80 |
| Contacto | 50 | 36 | 7 | 23 | 87 | 10 | 97 | 82 |
| | | | | | | | | |

**Tabla 1.** Comparación de métodos conglomerados respecto a la identificación de subestructuras genéticas bajo distintos modelos de migración que generan estructura genética de poblaciones.

+ Porcentaje de veces en que el K fue detectado, @ Porcentaje de individuos bien clasificados, * Porcentaje de nodos bien clasificados en la red. UPGMA (encadenamiento promedio) algoritmo jerárquico, K-means algoritmo no-jerárquico, SOM (self-organizing maps) algoritmo basado en redes neuronales y Bayesiano (Pritchard et al., 2000).

## 4. DISCUSIÓN

Los algoritmos de conglomerados confrontados en este trabajo funcionaron de forma diferencial en cada modelo biológico, para detectar los grupos o poblaciones, debido a la estructura genética explícita en cada uno. El método jerárquico para el modelo de islas identificó los 5 grupos simulados. Si bien Sagnard y sus compañeros (2002) sugieren usar un método de análisis discriminante (AD) en vez de algoritmos UPGMA para estudiar la estructura de diversidad genética y poder agrupar poblaciones geográficas o ecológicas, el AD requiere conocer a priori los grupos, mientras que el análisis de conglomerados jerárquicos se usa en la situación contraria, i.e. cuando no se conocen a priori la cantidad de grupos. El algoritmo K-means tuvo un buen desempeño en la detección de los grupos simulados bajo el modelo de islas. Fernández y sus compañeros (2007) lo citan como uno de los algoritmos divisivos más simples de calcular. Jombart y sus compañeros (2010) se refieren a k-means como el algoritmo natural a utilizar para definir el número de grupos cuando éste no se conoce a priori, debido a que utiliza el mismo modelo que el análisis discriminante y una medida similar en la diferenciación de grupos.

El algoritmo basado en redes SOM es una alternativa recientemente usada en estudios donde se desea encontrar la estructura genética en datos obtenidos a partir de marcadores moleculares. Fernández y Balzarini (2007) propusieron el uso de SOM junto a una aplicación desarrollada como herramienta para mejorar la visualización de los resultados en el contexto de abundante información genómica. Las adaptaciones específicas usadas en ese trabajo permitieron una mejor identificación de grupos de genes que el algoritmo K-means e incluso el algoritmo UPGMA. Sin embargo en nuestro estudio, con una cantidad sustancialmente menor de marcadores y sin el uso de la herramienta gráfica, el método SOM no fue el algoritmo que mejor identificó los grupos y las subestructuras emuladas. El método bayesiano basado en cadenas de Markov de Monte Carlo demostró un buen desempeño en el reconocimiento de la estructura simulada bajo el modelo de islas, pero no para el modelo de contacto, reduciendo su porcentaje de detección desde el 90% al 20%, respectivamente. Además, en la representación gráfica no se visualizan de manera clara los grupos, ni las subpoblaciones. Evanno y sus compañeros (2005) evaluaron el método bayesiano para detectar estructura genética sobre datos simulados bajo tres modelos de migración, concluyendo también que este algoritmo resulta eficaz no sólo para un modelo de isla, sino para esquemas más complejos de migración jerárquica.

## 5. COMENTARIOS FINALES

Para el modelo de islas el algoritmo que mejor identificó los grupos de subpoblaciones simulados fue el bayesiano. Para el modelo de contacto fue mejor el conglomerado jerárquico UPGMA, aunque la identificación de la estructura genética bajo este modelo fue pobre bajo todos los algoritmos. Los algoritmos SOM y bayesiano distinguieron mejor subpoblaciones que grupos de subpoblaciones. Los resultados sugieren que el desempeño de los algoritmos depende del modelo biológico subyacente para la estructuración genética de poblaciones, y que existe menos error de clasificación con el modelo de islas que con el de contacto. Nuevos desarrollos son necesarios para mejorar la identificación de la subestructura genética subyacente en situaciones biológicas

## BIBLIOGRAFÍA

Balloux, F. EASYPOP versión 1.7. (2001) A computer program for the simulation of population genetics. J. Heredity 92: 301-302.

Balzarini, M. y Di Rienzo, J. (2004) Info-Gen, Software para análisis estadístico genómico. Facultad de Ciencias Agropecuarias. Universidad Nacional de Córdoba. http://web.info-gen.com.ar. Versión actualizada 2010.

Chiappero, M., Panzetta-Dutari, G., Gómez, D., Castillo, E., Polop, J. y Gardenal, C. (2010) Contrasting genetic structure of urban and rural populations of the wild rodent Calomys musculinus (Cricetidae, Sigmodontinae). Mammalian Biology. doi:10.1016/j.mambio.2010.02.003.

Evanno, G., Regnaut, S. y Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14: 2611-2620.

Fernández, E. y Balzarini, M. (2007) Improving cluster visualization in self-organizing maps: Application in gene expression data analysis. Computers in Biology and Medicine 37:1677– 1689.

Fernández, E., Alvarez, M., Podhajcer, O., Stolovitzky, G. (2007) Genomica funcional: En busca de la función de los genes. National Academy of Science, Argentine. XIII 63-76.

Hedrick, P. (2005) Large variance in reproductive success and the Ne/N ratio. Evolution 59:1596-1599.

INFOSTAT. Manual del usuario. (2008) Grupo InfoStat, FCA, Universidad Nacional de Córdoba. Primera Edición, Editorial Brujas. Córdoba, Argentina.

Jombart, T., Devillard, S. y Balloux, F. (2010) Discriminant analysis of principal compo-

nents: a new method for the analysis of genetically structured populations. BMC Genetics 11:94.

Kohonen, T. (1997) Self-Organizing Maps. Second Ed. Springer. Berlin.

Milligan, G. (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45:325-342.

Pritchard, J., Stephens, M. y Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Sagnard, F., Barberot, C. y Fady, B. (2000) Structure of genetic diversity in Abies alba Mill. from southwestern Alps: multivariate analysis of adaptive and non.adaptive traits for conservation in France. Forest Ecology and Management 157:175-189.

Sokal, R. y Michener, C. (1958) A statistical methods for evaluating systematic relationships. University of Kansas Science Bulletin 38:1409-1438.

Wright, S. (1931) Evolution in mendelian populations. Genetics 16:97-159.



| Referencia | Fecha de recepción | Fecha de aprobación |
|---|---|---|
| Peña Malavera Andrea Natalia,<br>Bruno Cecilia, Teich Ingrid y<br>Balzarini Mónica. Clasificación en la identificación de estructura genética de poblaciones a partir de datos moleculares.<br>Revista *Tumbaga* (2010), 5, 225 - 236 | Día/mes/año<br>30/08/2010 | Día/mes/año<br>1/09/2010 |

# REVISTA CIENTÍFICA TUMBAGA

## "CIENCIA EN CONSTRUCCIÓN"

## POLÍTICA EDITORIAL

1. Se publicarán los siguientes tipos de documentos:
   - **Artículos de investigación científica y tecnológica** que presenten, de manera detallada, los resultados originales de proyectos de investigación.
   - **Artículos de reflexión** que presenten resultados de investigación desde una perspectiva analítica, interpretativa o crítica del autor, sobre un tema específico, recurriendo a fuentes originales.
   - **Artículos de revisión** elaborados con base en una investigación donde se analicen, sistematicen e integren los resultados de investigaciones publicadas o no publicadas, sobre un campo en ciencia o tecnología, con el fin de dar cuenta de los avances y las tendencias de desarrollo. Deberán caracterizarse por presentar una cuidadosa revisión bibliográfica de por lo menos 50 referencias.

   También se publicarán **Artículos cortos**, **Reportes de caso**, **Revisiones de temas**, **Cartas al editor**, **Traducciones**, **Documentos de reflexión no derivados de investigaciones, Reseñas bibliográficas** (de acuerdo con la definición de COLCIENCIAS, disponible en www.colciencias.gov.co), **y resúmenes de trabajos de grado y de tesis de postgrado.** Los trabajos deben estar relacionados con las áreas de biología, química, física, matemáticas, estadística. También se incluirán artículos de divulgación dirigidos a una comunidad no especializada.

2. Se recibirán artículos escritos en español, inglés y portugués.

   Los títulos de los artículos deberán Escribirse en español e inglés.

3. Los trabajos presentados deben ser originales e inéditos, y no deben estar siendo evaluados en otras revistas. El Comité Editorial considerará la re publicación de traducciones o artículos cuando éstos sean de gran relevancia y hayan sido publicados en revistas de difícil acceso en el medio.

4. La extensión máxima de los trabajos debe ser de 24 páginas en tamaño carta (8.5 x 11.0 pulgadas), a doble espacio, letra Arial 12, con márgenes izquierdo y derecho de 3 cm. Cuando se trate de artículos de revisión la extensión máxima será de 44 páginas.

5. Los artículos incluirán los siguientes apartes: título, autores (apellido, nombre, institución, correo electrónico del autor para correspondencia); palabras o frases clave (cinco según las categorías de la Nomenclatura Internacional de la UNESCO); resumen (en el idioma original y en inglés, con una máxima extensión de 250 palabras); introducción; desarrollo (en caso de un artículo de investigación incluirá metodología, resultados y análisis de los mismos); conclusiones y bibliografía (según las normas APA). Las notas al artículo se presentarán al final del mismo y antes de la bibliografía.

   La traducción del resumen debe ser realizada por una persona competente en relación con el manejo de la lengua, y en todos los casos los autores deben evitar recurrir a traductores automáticos.

6. Los artículos deben ser enviados en formatos *doc* o *tex*, cuando incluyan expresiones matemáticas. En los casos en los que los autores estén interesados en remitir artículos en formato *tex* deberán ponerse en contacto con el Comité Editorial con el propósito de que se les remita la plantilla correspondiente.

7. Cuando los artículos contengan imágenes éstas deberán enviarse por separado, en una resolución mínima de 300 dpi, y en formatos jpg, bmp o tiff, indicando en el nombre de cada archivo el título que le corresponda a la imagen en el artículo. El archivo de texto correspondiente al artículo podrá contener o no las imágenes, siempre y cuando en él se indique claramente su ubicación, y se incluyan las respectivas leyendas o pies de gráfica. Las tablas y gráficos deberán remitirse en extensiones xls o doc. Para las estructuras químicas debe utilizarse *Acdlabs* o *Chemdraw*.

8. Todos los artículos serán sometidos a una revisión preliminar por parte del Comité Editorial, en la que se analizará la adecuación a las políticas de publicación, y posteriormente serán remitidos a pares académicos para su evaluación según los siguientes criterios: originalidad, claridad y coherencia, interés, pertinencia e importancia del tema tratado, nivel científico -nivel de conceptualización-, y aportes a la formación científica.

9. Los artículos podrán ser sometidos a un comité de ética según sea pertinente.

10. Los artículos se recibirán permanentemente en la dirección electrónica *tumbaga@ut.edu.co*, o en la Facultad de Ciencias de la Universidad del Tolima (Barrio Santa Elena, Ibagué, Tolima). El autor deberá diligenciar los formatos anexos para el envío del artículo y enviar el archivo correspondiente sin incluir en él sus datos personales, para facilitar el trámite de evaluación correspondiente.

11. La Revista se distribuirá a las bibliotecas de las 4 Universidades afiliadas a Ascun, Asofacien y Ascofade. Igualmente, estará disponible para investigadores, profesores de los diferentes niveles educativos, estudiantes universitarios y la comunidad en general. Los interesados en suscribirse pueden diligenciar el formato anexo.

Estamos seguros de que la difusión del producto de sus investigaciones a través de nuestra revista ofrecerá grandes satisfacciones tanto personales como académicas para nuestra comunidad. Por ello, lo invitamos a hacer parte activa de su desarrollo.

Un cordial saludo,

**Comité Editorial de la Revista**
tumbaga@ut.edu.co