ANÁLISIS DE DATOS MULTIVARIANTES CON COORDENADAS PARALELAS

Dr. Estadística Multivariante Aplicada

carlomagnocr@gmail.com

Resumen

En este artículo, se estudian las Coordenadas Paralelas (Coords ||), que es un sistema de visualización que permite representar *n-dimensiones* en un sistema bidimensional. En este sistema, cada eje vertical (ordenada) representa un atributo (dimensión) que puede ser continuo o categórico. Cada uno de los ejes verticales de un sistema de Coords || puede tener su propia escala o definirse todos con una sola escala, la primera forma nos permite la visualización de híper-superficies y el análisis del funcionamiento del conjunto de datos, con la segunda podemos hacer un análisis de las relaciones entre las variables.

En general, las Coords || son una técnica de visualización donde las dimensiones son simbolizadas como una serie de ejes paralelos perpendiculares, con la misma separación entre ellos (equidistantes) y donde los valores están representados. Cada eje representa una coordenada en la dimensión correspondiente. Uniendo con líneas los ejes, podemos simbolizar los puntos en ***n***-dimensiones. Asimismo, un punto en un espacio ***n***-dimensional es transformado en una línea poligonal a través de n ejes paralelos como ***n-1*** segmentos de línea.

El objetivo del trabajo es analizar las coordenadas paralelas como técnica de análisis multivariante y así aprovechar su potencial para la visualización n-dimensional de los datos.

Palablas claves: coordenadas paralelas, visualización multidimensional de datos.

Abstract

In this paper, we study the Parallel Coordinates (Coords ||), a display system that allows us to represent n-dimensions on a two-dimensional system. In this system, each vertical axis (ordinate) represents an attribute (dimension) that can be continuous or categorical. Each of the vertical axis Coords || system may have its own scale or defined all with a single scale, the first form allows the display of hyper-surfaces and performance analysis of the data set, the second can to analyze the relationships between variables.

In general, Coords || are visualization techniques where the dimensions are symbolized as a series of parallel lines perpendicular to the same separation between them (equidistant) and where the values are represented. Each axis represents a coordinate in dimension. Lines joining the axes, we can symbolize the points in **n**-dimensions. Also, a point in a **n**-dimensional space is transformed into a polygonal line through n parallel axes and **n-1** line segments.

The study aims to analyze the parallel coordinates as multivariate analysis technique and exploit their potential for n-dimensional visualization of data.

Keywords: parallel coordinate, multidimensional data visualization.

INTRODUCCIÓN

En los últimos años, se han propuesto gran cantidad de métodos gráficos para el Análisis Exploratorio de Datos Espaciales (*AEDE*) aunque, existen pocos estudios que valoren la utilidad y efectividad de todos ellos. Podría afirmarse que un buen método gráfico de *AEDE* es aquél capaz de analizar y representar características en toda su distribución espacial: variabilidad, tendencia central, clúster y puntos atípicos.

Entre los muchos gráficos propuestos por el *AED* clásico para el análisis multivariante, estudiaremos los gráficos de *coordenadas paralelas (*Coords ||*).* Estas fueron propuestas por Alfred Inselberg de la Universidad de Illinois en 1959, como metodología para la visualización de *n-dimensiones* en problemas de datos multivariantes.

Las Coords || es un sistema de visualización que permite representar *n-dimensiones* en un sistema bidimensional. En este sistema, cada eje vertical (ordenada) representa un atributo (dimensión) que puede ser continuo o categórico. Cada uno de los ejes verticales de un sistema de Coords || puede tener su propia escala o definirse todos con una sola escala, la primera forma nos permite la visualización de híper-superficies y el análisis del funcionamiento del conjunto de datos, con la segunda podemos hacer un análisis de las relaciones entre las variables.

En general, las Coords ‖ son una técnica de visualización donde las dimensiones son simbolizadas como una serie de ejes paralelos perpendiculares, con la misma separación entre ellos (equidistantes) y donde los valores están representados. Cada eje representa una coordenada en la dimensión correspondiente. Uniendo con líneas los ejes, podemos simbolizar los puntos en *n*-dimensiones. Asimismo, un punto en un espacio *n*-dimensional es transformado en una línea poligonal a través de n ejes paralelos como $n-1$ segmentos de línea.

De tal forma, el vector $x = [x_1, x_2, \cdots, x_6]$ es representado por medio de x_j en la coordenada 1, x_2 en la coordenada 2 y así sucesivamente, hasta la x_n en al coordenada *n*. A partir de la representación resultante, podemos sacar conclusiones al respecto, por ejemplo sobre la relación entre las variables. Un grupo de líneas proyectas bastante próximas una con otra nos indicará un grado de asociación positiva entre las variables que la componen.

En Coords ‖ pueden visualizarse muchas dimensiones, limitadas en la práctica por la resolución horizontal de la pantalla del ordenador. A medida que el número de dimensiones aumenta, las coordenadas tendrán que ser representadas muy próximas unas con otras generando mayores dificultades para la percepción de los patrones. Tienen la ventaja que nos permite visualizar una cantidad mayor de variables y sus relaciones, que no es posible con los métodos tradicionales (2 ó 3 dimensiones).

El objetivo del trabajo es analizar las coordenadas paralelas como técnica de análisis multivariante y así aprovechar su potencial para la visualización n-dimensional de los datos.

EL PLANO $\mathbb{R}^2$

El sistema de coordenadas cartesianas se construye por ejes en el plano recto mutuamente perpendiculares que se cortan en el origen y cada punto del plano es definido mediante dos números (a, b). En tanto, en $\text{Coords} \parallel$ los ejes son paralelos y el número (a, b) se representa utilizando una *línea* que conecta los valores de a en la abscisa x_1 y b en el eje x_2 (*Figura 1*).

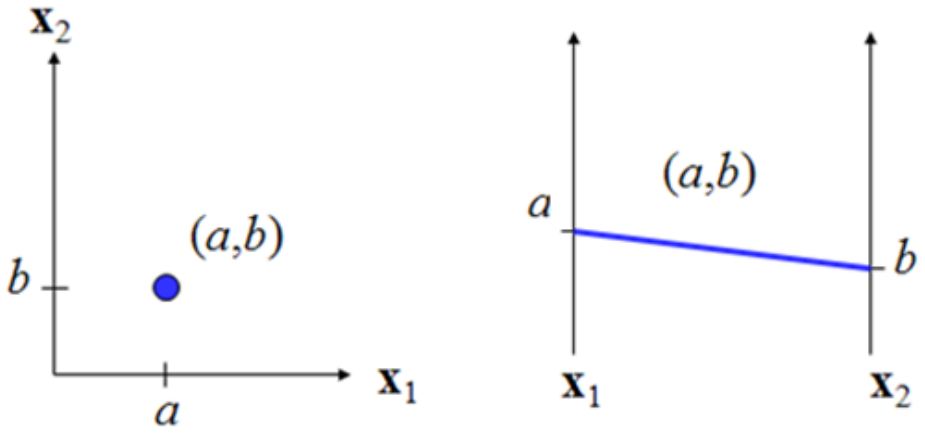

Figura 1. Dualidad punto – línea en coordenadas cartesianas y paralelas.

Un punto N-dimensional $P = (p_1, \cdots, p_N)$ es representado por una línea *poligonal*[1] con $\bar{P}$ vértices para los valores p_i donde $i = 1, \cdots, N$. Considere

[1] *Línea poligonal* es aquella formada solo por segmentos de recta unidos.

la representación de un punto en seis dimensiones definido como un vector *p*, donde $p = [p_1, , \cdots, p_6]$ toma los valores $P = (5, -5, 10, 15, 5, -10)$. La línea N-dimensional ℓ se dibuja como,

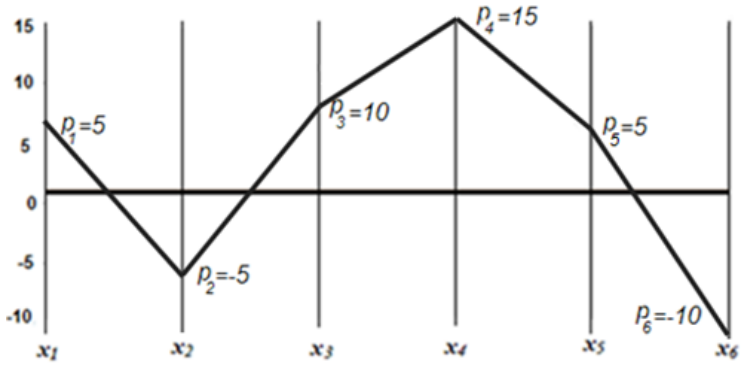

Figura 2. Visualización de un punto n-dimensional $\mathrm{Coords}\parallel$.

La línea ℓ en $\mathbb{R}^N$ se expresa con cierta manipulación algebraica, en términos de $N-1$ hiperplanos paralelos. Equivalentemente, este conjunto de puntos (descritos por las N-tuplas) satisfacen un conjunto de $N-1$ ecuaciones lineales independientes,

$$\ell_{ij} = x_i = m_{ij}x_j + b_{ij}$$

La ecuación describe la proyección de ℓ sobre el plano bidimensional $x_i x_j$. Más específicamente, las N-tuplas se representan de la siguiente manera:

$$\ell: \begin{cases} \ell_{1,2}: & x_2 = m_2 x_1 + b_2 \\ \ell_{2,3}: & x_3 = m_3 x_2 + b_3 \\ & \cdots\ \cdots\ \cdots \\ \ell_{i-1,i}: & x_i = m_i x_{i-1} + b_i \\ & \cdots\ \cdots\ \cdots \\ \ell_{N-1,N}: & x_N = m_N x_{N-1} + b_N \end{cases}$$

Cada ecuación contiene un par de variables *adyacentes*. En el plano $x_{i-1}x_i$ la relación es indicada por la línea $\ell_{i-1,i}$ y por la correspondencia $punto \leftrightarrow linea$ es representada como (Inselberg, 1992). Un caso en particular, es la representación de una recta en tres dimensiones (*Figura 3*). El *hiperplano*[2] se visualiza en Coords ‖ utilizando el programa MATLAB (A*nexo 1*).

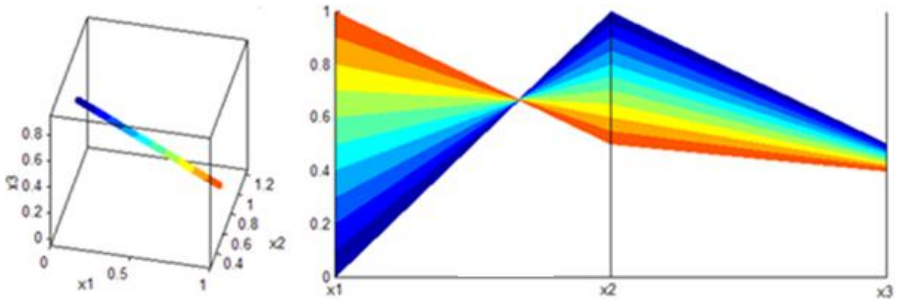

Figura 3. Visualización de una recta tridimensional en Coords ‖.

VISUALIZACIÓN DE RESULTADOS

La exploración visual de datos multidimensionales es de gran interés en estadística y en la visualización de información. Ayuda a encontrar tendencias y relaciones entre dimensiones. Al visualizar datos multidimensionales, cada variable puede trazar cierta entidad o cualidad gráfica.

Una apropiada visualización: revela claramente la estructura dentro de los datos, puede ayudar a identificar mejor patrones (modelos) y permiten

[2] Un *hiperplano*, es la generalización al espacio n-dimensional del concepto de recta.

detectar afloramientos. El *alboroto visual*[3], está caracterizado por las entidades visuales apretadas y desordenadas que distorsionan la estructura en las representaciones visuales, que dificultan la identificación rápida de información relevante. El desorden es contrario a la estructura visual; corresponde a todos los factores que interfieren con el proceso de encontrar estructuras y obstaculiza comprensión del contenido de las exhibiciones.

Las correlaciones entre las dimensiones pueden ser descubiertas concentrándose en las intersecciones de las polilíneas, al detectar grupos de observaciones con pendientes comunes en las líneas de conexión inter-variables, poniendo de relieve un determinado tipo de correlación entre dichas variables (positiva, negativa o nula). Cuando la correlación lineal simple entre dos variables tiende a uno, tenemos en $\mathrm{Coords}\parallel$ "*efecto del cruce*" (Wegman, 1990). Así la estructura de la correlación se puede diagnosticar fácilmente; esta configuración la representó Griffen (1958) y la utilizó como dispositivo gráfico para computar la Tau de Kendall. La fórmula de cálculo que esbozó fue la siguiente,

$$r = 1 - \frac{4X}{n(n-1)}$$

donde *X* es el número de intersecciones de líneas resultantes de la conexión de dos variables en $\mathrm{Coords}\parallel$. El número de empalmes es invariante a cualquier transformación monótona de *x* o de *y* en

[3] El *desorden visual* es un problema en todo tipo de diseño que cumpla alguna función como lo son los gráficos o bien las interfaces interactivas (software o páginas web por ejemplo).

coordenadas cartesianas. Si hay una relación lineal perfecta positiva, sin encuentros entre líneas, entonces *X=0* y *r=1.* La Figura 4 muestra el axioma:

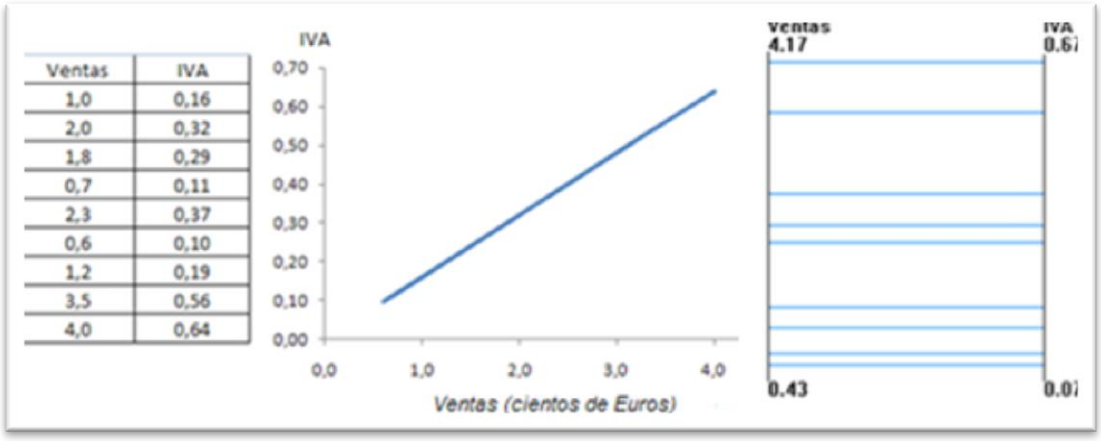

Figura 4. Correlación lineal positiva perfecta en coordenadas cartesianas y Coords ‖.

Igualmente, cuando se tiene una correlación perfecta negativa todas líneas se intersecan (*Figura 5*). El número de intersecciones es $\binom{n}{2}$, por tanto la fórmula de cálculo es,

$$r = 1 - \frac{4\binom{n}{2}}{n(n-1)} = -1$$

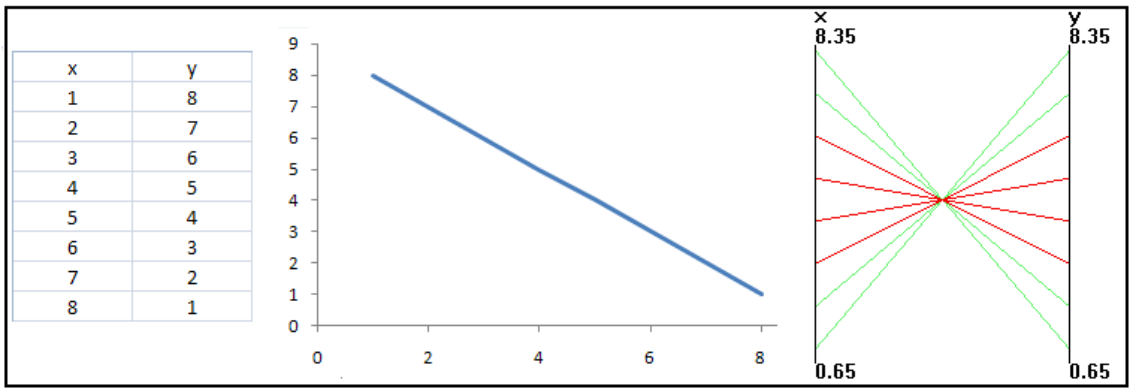

Figura 5. Correlación negativa perfecta en coordenadas cartesianas y $\mathrm{Coords}\parallel$.

Una de las etapas en un análisis multivariante es la detección de valores extremos; que son observaciones alejadas de centro de gravedad de los datos. La presencia de valores extremos produce modificaciones arbitrarias en los valores de los estimadores de máxima verosimilitud y por consecuencia en los resultados (o conclusiones).

Los valores extremos pueden identificarse desde una perspectiva univariante, bivariante o multivariante. Específicamente, la evaluación multivariante implica una evaluación de cada observación a lo largo de un vector de variables. Se puede utilizar una distancia Euclídea para determinar si una observación es un valor extremo, pero esta no es eficiente cuando existe dependencia entre las observaciones ya que no considera la estructura de correlación.

Para evitar estos problemas podemos representar el conjunto de datos multivariantes sobre $\mathrm{Coords}\parallel$, determinando así las direcciones de proyección de los valores extremos. Es recomendable estandarizar las variables para poder compáralas y permitir un mejor descubrimiento del patrón anormal (Novotny & Hauser, 2006). La *Figura 6* permite visualizar un valor extremo multidimensional; el valor se representa por una polilínea a través de las abscisas en la parte inferior de la figura (en rojo).

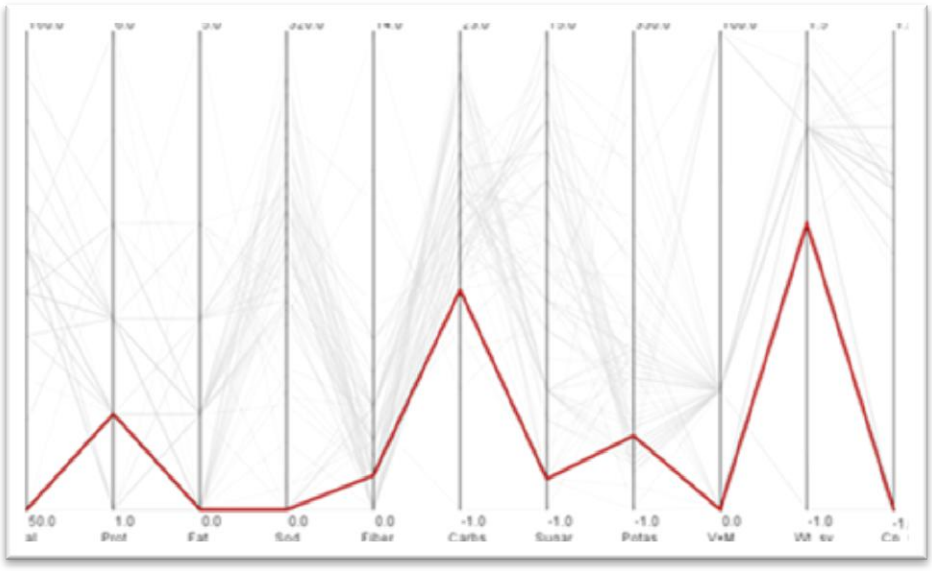

Figura 6. Representación de un valor extremo multidimensional en Coords ||.

Las Coords || resultan útiles para captar agrupamientos (“clústeres”) entre observaciones cuando sus correspondientes líneas presenten una forma similar (por ejemplo, estén agrupadas de forma diferente en el gráfico). El descubrimiento de grupos o racimos de polilíneas[4] diferenciadas del resto se consigue cambiando los órdenes de las dimensiones, para procurar que las relaciones de los datos puedan ser visualizadas. La visualización en Coords || de grupos, es una cuestión de percepción y a la vez tratar de descubrir las relaciones entre las coordenadas (variables), utilizando diferentes reordenamientos de los ejes que nos permitan sacar conclusiones válidas.

[4] La representación mediante *polilíneas* está basada en la definición de un objeto mediante *n-1* segmentos rectos consecutivos determinados por una lista de *n* puntos, cuyos puntos de inflexión denominamos vértices. Una polilínea puede ser cerrada, constituyendo de esta forma un polígono.

Las polilíneas que tienden a estar cercanas constituirán un grupo a diferencia de aquellas que se separan y cuando hay líneas que no pertenecen a ningún grupo (fuera de los patrones) pueden considerarse como valores extremos (Chou et al.,1999). Como ejemplo, se presenta la visualización de dos clúster generados en tres dimensiones y en Coords ‖ con un pequeño programa elaborado en sistema Matlab[5] (A*nexo 2*).

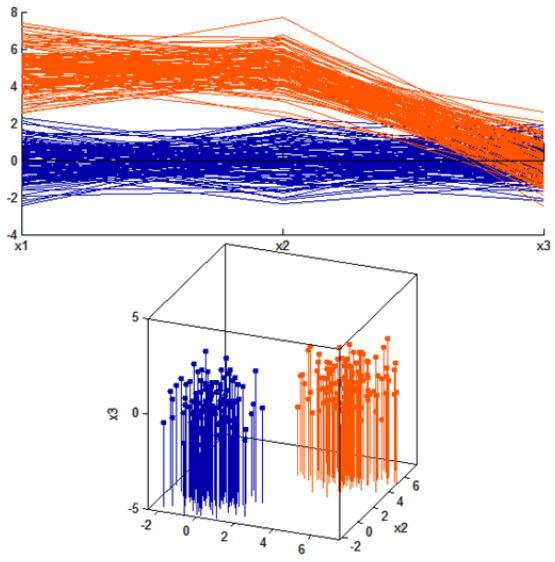

Figura 7. Visualización de clúster en 3D.

APLICACIONES Y RESULTADOS

En esta sección se describe una aplicación que hemos realizado usando Coords ‖, con el propósito de poner en práctica el marco teórico visto en anteriormente en el artículo.

[5] **MATLAB** es la abreviatura de Matrix Laboratory (laboratorio de matrices). Es un programa de análisis numérico creado por The MathWorks en 1984.

FISHER (1936) clasificó tres especies del género de flores de Iris (*Setosa, Versicolor y Virginic*a, claramente diferenciadas entre sí por el color), con una muestra total de 150 plantas y cuatro medidas asociadas al sépalo de la flor: longitud del sépalo, ancho del sépalo, longitud del pétalo y ancho del pétalo (en centímetros).

En la *Figura 4.1*, se visualiza una diferencia en el comportamiento de las cuatro medidas asociadas al sépalo de la flor según las especies. Los tres géneros de flores Iris forman clúster perfectamente diferenciados y en especial, las medidas de la setosa son muy diferentes respecto a versicolor y virginica.

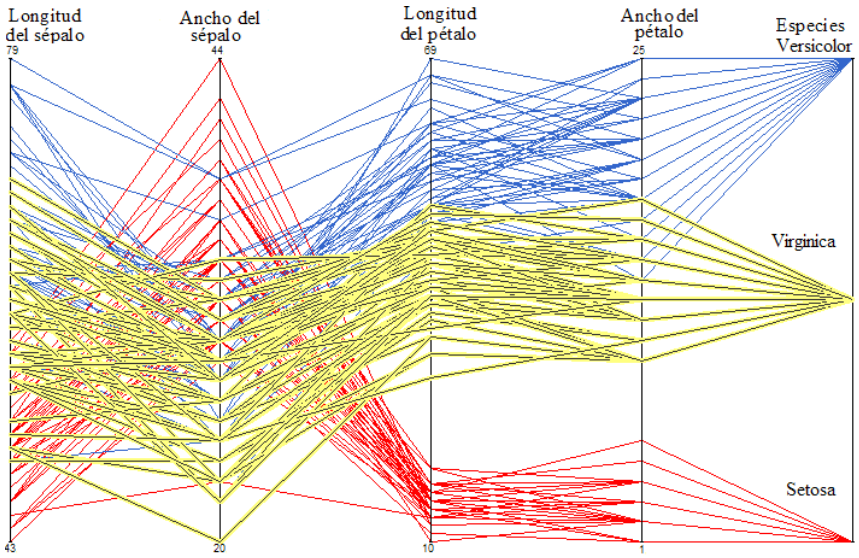

Figura 8. Representación en Coords || de los datos de IRIS, Fisher (1936).

La correlación entre las medidas está asociada con el género de flores. Se observa correlaciones muy diferentes entre el ancho del sépalo respecto a la longitud del sépalo y del pétalo en la especie setosa. Sin embargo, la asociación es similar entre la longitud del sépalo y el ancho del pétalo indiferentemente de la especie.

La variedad setosa presenta una variabilidad pequeña en la longitud y el ancho del pétalo; aumenta su variancia en forma significativa respecto al ancho del sépalo. Los modelos matemáticos que siguen las medidas son muy diferentes dependiendo de las especies, especialmente en las medidas de longitud del sépalo, ancho del sépalo y ancho del pétalo, con una función estrictamente cóncava en las variedades virginica y versicolor; en tanto una función estrictamente convexa en setosa.

CONCLUSIONES

Hemos mostrado que las Coords ‖ son un método novedoso de visualizar de un plano n-dimensional en una representación de dos dimensiones y que son especialmente útil para mostrar patrones en los datos o para percibir relaciones entre variables. Las Coords ‖ tienen la ventaja respecto a otros métodos de representación, como el Biplot, que son: fáciles de construir, no se requieren conocimientos avanzados de matemáticas. Además, proporcionan un medio para poder, se observan patrones o las tendencias de las variables para los diagnósticos de modelos matemáticos y permiten observar el grado de correlación entre pares de

variables colindantes. También, es una herramienta que nos ayuda a detectar los valores extremos multivariantes presentes en conjunto de datos y puede utilizarse para la visualización de clúster. Aunque su potencialidad para el análisis de datos, se máxima cuando la visualización es interactiva (o dinámica).

BIBLIOGRAFÍA

Andrienko, G.; Andrienko, N. (2001). Constructing Parallel Coordinates Plot for Problem Solving. In Proc. 1st International Symposium on Smart Graphics, March 21-23, 9–14.

Anselin, L. (1999). The Future of Spatial Analysis in the Social Sciences. Geographic Information Sciences, 5(2):67-76.

Artero, A.O.; Ferreira, O.M.; Levkowitz, H. (2004). Uncovering Clusters in Crowded Parallel Coordinates Visualizations. IEEE Symposium on Information Visualization, October, 81–88.

Chou, S.Y.; Lin, S.W.; Yeh, C.S. (1999). Cluster Identification with Parallel Coordinates. Elsevier Science, 20(6):565-572.

Earnshaw, K.W.; Brodlie, L.A. (1992). Scientific Visualization: Techniques and Applications. Carpenter et al., editors. Springer-Verlag.

Fua, Y.H.; Ward, M.O.; Elke E.A. (1999). Hierarchical Parallel Coordinates for Exploration of Large Datasets. In Proceedings of the Conference on Visualization, 43–50.

Graham, M.; Kennedy, J. (2003). Using Curves to Enhance Parallel Coordinate Visualizations. IEEE Computer Society, Proceedings of the Seventh International Conference on Information Visualization, 10-16.

Griffen, H.D. (1958). Graphic Computation of Tau as a Coefficient of Disarray. American Statistical Association, 53(282):441-447.

Haining, R.S.W.; Signoretta, P. (2000). Providing Scientific Visualization for Spatial Data Analysis: Criteria and an Assessment of SAGE. Journal of Geographical Systems, 2:121-140.

Hauser, H.; Ledermann, F.; Doleisch, H. (2002). Angular Brushing of Extended Parallel Coordinates. Proceedings of IEEE Symposium on Information Visualization 2002, Boston, Massachusetts, Oct. 2002, IEEE Computer Society, 127–130.

Hauser, H.; Novotry, M. (2006) Outlier - Preserving Focus + Context Visualization in Parallel Coordinates. IEEE Transactions on Visualization and Computer Graphics, 12(5):893-900.

Inselberg, A. (1992). The Plane R^2 with Coordinate Parallel. Computer Science and Applied Mathematics Departments. Tel Aviv, University, Israel.

Inselberg, A.; Dimsdale, B. (1990). Parallel Coordinates: a tool for Visualizing Multi-Dimensional Geometry. Proceedings of the First IEEE Conference on Visualization, 361-378.

Izhakian, Z. (2004). New Visualization of Surfaces in Parallel Coordinates – Eliminating Ambiguity and Some “Over-Plotting”. Journal of WSCG, 1-3(12):183-191.

Streit, M.; Ecker, C.R. Osterreicher, K.; Steiner, E.G.; Bischof, H.; Bangert, C.; Kopp, T.; Rogojanu, R. (2006). 3D Parallel Coordinate systems – A New Data Visualization Method in the Context of Microscopy –based Multicolor Tissue Cytometry. 69(7):601–611.

Tukey, J. W. (1977). Exploratory Data Analysis. Reading: Addison-Wesley.

Wegman, E. (1990). Hyperdimensional Data Analysis Using Parallel Coordinates. Journal American Statistical Association, 85(411):664-675.

ANEXO 1:
LÍNEA 3D EN COORDENADAS PARALELAS

```
clf
clear all
npoints=300;
t=linspace(0,1,npoints)
position(1,:)= t;  %x
position(2,:)= 1-.5*t ;  %y
position(3,:)= 0.5-.1*t;
s=t;
map=jet(10) ;
s = 10*(s - min(s))/(max(s)-min(s))+1;
subplot(2,1,1)
hold on
for i = 1:npoints
    cindex = min( fix(s(i)), 10);
    plot(position(:,i), 'color', map(cindex,:) ); end
axisx = [1,3; 1,1; 2,2; 3,3; ]';
axisy = [0,0; 0,1; 0,1; 0,1; ]';
line(axisx, axisy, 'color', 'k');
set(gca, 'xticklabel', {'x1', ' ', 'x2',' ','x3'})
subplot(2,1,2)
Title('Line: x1=t;  x2=1-.5*t;  x3=0.5-.1*t;')
set(gca,'color','white')
box on
view(22,63)
axis vis3d
xlabel('x1');ylabel('x2');zlabel('x3');
hold on
for i = 1:npoints
    cindex = min( fix(s(i)), 10);
    plot3(position(1,i),position(2,i),position(3,i),...
        'markerfacecolor', map(cindex,:),...
        'markersize',2 ,'color',map(cindex,:),...
        'marker','o',...
        'linewidth',2)
  end
  rotate3d on
```

ANEXO 2:
CLUSTER 3D EN COORDENADAS PARALELAS

```
clf
clear all
npoints=250;
for i = 1:npoints/2;
   position(1,i)= randn;
   position(2,i)= randn;
   position(3,i)= randn;
   s(i)=1;end
delta=5;
for i = npoints/2+1:npoints;
   position(1,i)= randn+delta;
   position(2,i)= randn+delta,
   position(3,i)= randn ;
    s(i)=2;end
map=jet(10);
s = 10*(s - min(s))/(max(s)-min(s))+1;
subplot(2,1,1)
hold on
for i = 1:npoints
    cindex = min( fix(s(i)), 10);
    plot(position(:,i), 'color', map(cindex,:) ); end
axisx = [1,3; 1,1; 2,2; 3,3]';
axisy = [0,0; 0,1; 0,1; 0,1]';
line(axisx, axisy, 'color', 'k');
set(gca, 'xtick', [1 2 3]);
set(gca, 'xticklabel', {'x1', 'x2','x3'})
subplot(2,1,2); set(gca,'color',[1 1 1])
box on; axis vis3d
view(22,23)
xlabel('x1'); ylabel('x2'); zlabel('x3');
hold on
for i = 1:npoints
    cindex = min( fix(s(i)), 10);
    plot3(position(1,i),position(2,i),position(3,i),...
        'markerfacecolor', map(cindex,:),...
        'markersize',2 ,'color',map(cindex,:),...
        'marker','o',...
        'linewidth',2)
    line ( [position(1,i) position(1,i)],...
        [position(2,i) position(2,i)],...
        [position(3,i), -5] ,'color',map(cindex,:));end
set(gca,'zlim',[-5 5])
rotate3d on
```