Pensamiento Actual. Universidad de Costa Rica.
Volumen 11 - No. 16-17, 2011 ◆ ISSN 2215-3586 ◆ 81 - 91

Análisis de datos multivariantes con coordenadas paralelas

Carlomagno Araya Alpizar[1]
Recibido: 5 de mayo de 2011 / Aprobado: 27 de setiembre de 2012

Resumen

En este artículo, se estudian las Coordenadas Paralelas (Coords ||), que son 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 comportamiento 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, con la misma separación entre ellos (equidistantes) y donde los valores son 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 datos.
Palabras claves: coordenadas paralelas, visualización multidimensional de datos.

Abstract

In this paper, Parallel Coordinates Coords ||, are studied, a display system that allows us to represent n-dimensions in two-dimensions. In this system, each vertical axis (ordinate) represents an attribute (dimension) that may be either continuous or categorical. Each of the vertical axes of a Coords || system may have its own scale or define everything within a single scale. The first form allows us the display of hyper-surfaces and performance analysis of the data set, while the second allows us to analyze the relationships between variables.

In general terms, Coords ||are a visualization technique where the dimensions are symbolized by a series of parallel and thus providing a space where the values are represented. Each axis represents a coordinate in the corresponding dimension. If the axes are linked by lines, the points in n-dimensions are symbolized. Likewise, a point in a n-dimensional space is transformed into a polygonal line through n parallel axes and $n-1$ line segments.

The aim of this study is to analyze the parallel coordinates as a multivariate analysis technique and thus 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. De acuerdo con lo anterior, podría afirmarse que un buen método gráfico de *AEDE* es aquel capaz de

[1]Doctor en Estadística Multivariante Aplicada, Profesor de Estadística, Universidad de Costa Rica, Sede de Occidente, carlomagnocr@gmail.com

analizar y representar características en toda su distribución espacial: variabilidad, tendencia central, clúster y puntos atípicos.

Entre los muchos gráficos propuestos por el *AEDE* 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 es 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 simbolizado 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 permiten 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 Coords ‖ 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*).

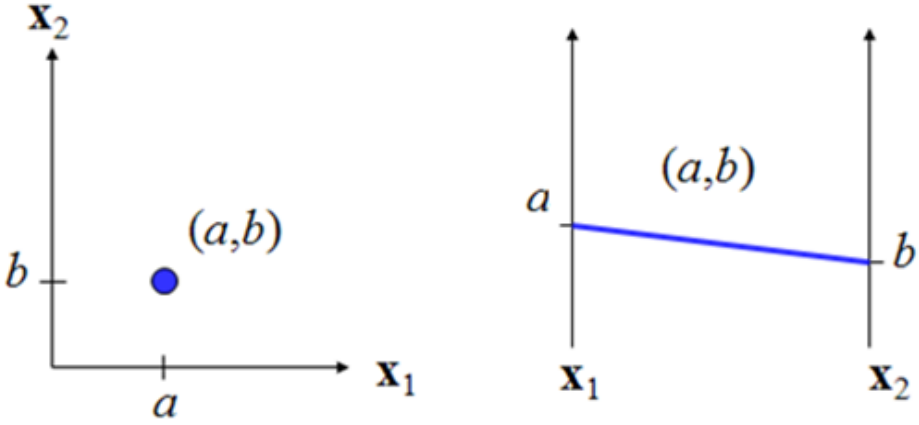

Figura 1. Representación de un punto (a, b) en coordenadas cartesianas y paralelas.

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

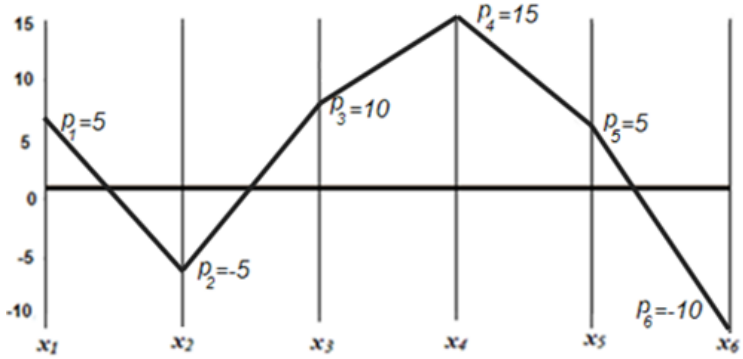

Figura 2. Visualización de un punto de 6-dimensional en Coords ‖.

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:

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

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

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:

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

Cada ecuación contiene un par de variables *adyacentes*. Un caso en particular, es la representación de una recta en tres dimensiones (*Figura 3*). Las unidades de medidas de x_1, x_2 y x_3 pueden ser cualquiera y solamente modifican la trayectoria de la recta.

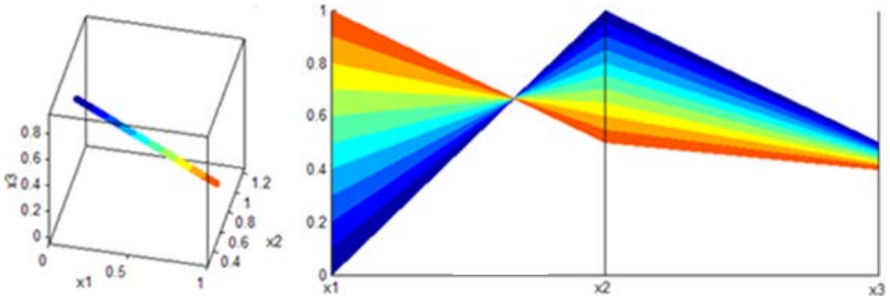

Figura 3. Visualización de una recta tridimensional en coordenadas cartesianas (izquierda) y Coords ‖ (derecha)

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 permite detectar afloramientos. El *alboroto visual*[2], 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

[2] 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).

desorden es contrario a la estructura visual; corresponde a todos los factores que interfieren con el proceso de encontrar estructuras y obstaculiza la 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 Coords ‖ "*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 Coords ‖. El número de empalmes es invariante a cualquier transformación monótona de *x* o de *y* en 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:

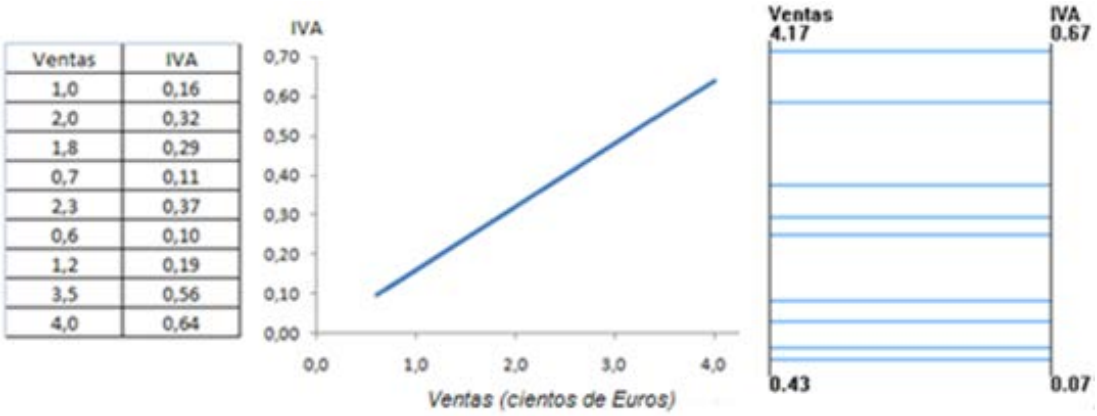

Figura 4. Visualización de la correlación positiva entre el monto de las ventas e impuestos de ventas (en miles de colones) en coordenadas cartesianas y Coords ‖.

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,

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

Pensamiento Actual. Volumen 11 - No. 16-17, 2011. Universidad de Costa Rica

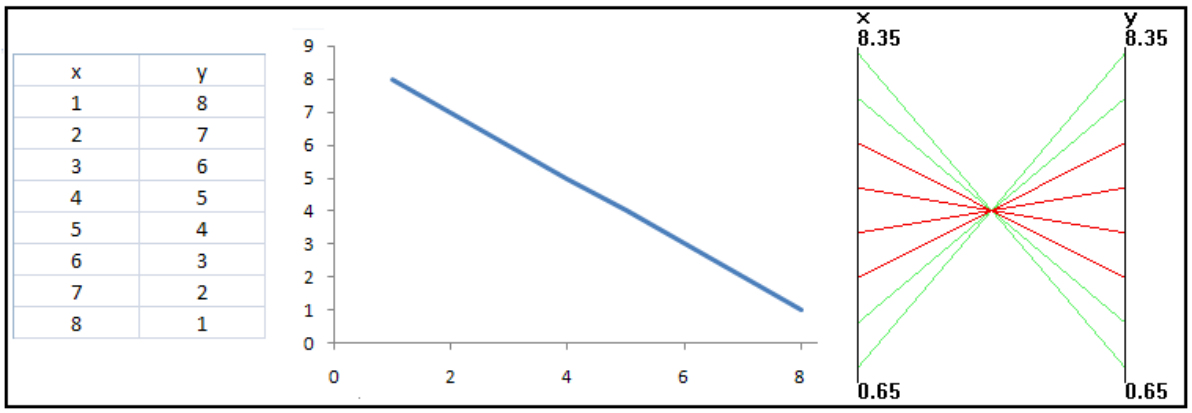

Figura 5. Representación de la correlación negativa entre dos variables (x, y) en coordenadas cartesianas y Coords ‖.

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 Coords ‖, determinando así las direcciones de proyección de los valores extremos. Es recomendable estandarizar las variables para poder compararlas 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).

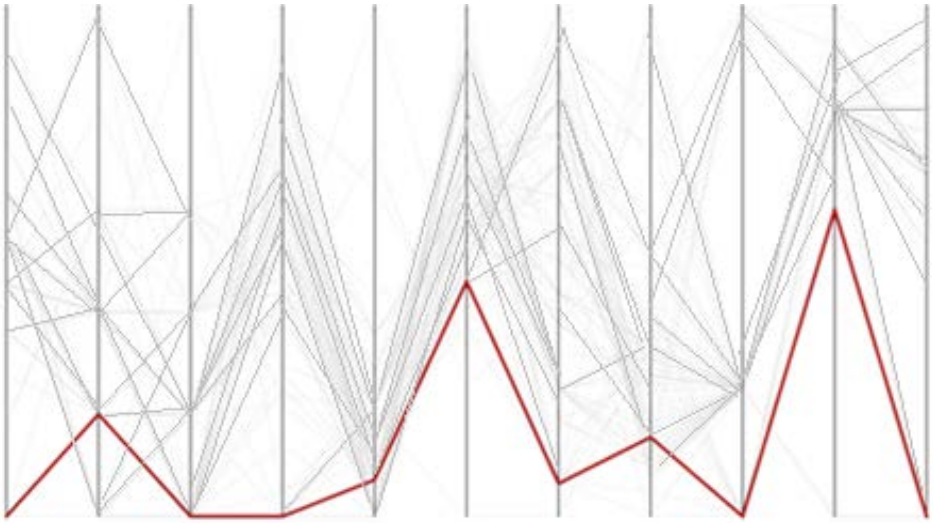

Figura 6. Representación de un valor extremo en 11-dimensional en Coords ||.

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[3] 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.

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).

[3] La representación mediante *polilíneas* está basada en la definición de un objeto mediante *n-1* segmentos rectos consecutivos determinados por una lista de

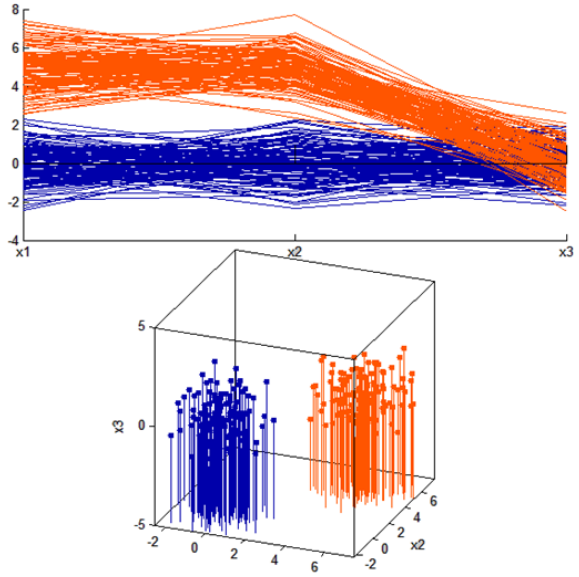

Figura 7. Visualización de dos grupos (o clúster) en tres dimensiones (x_1, x_2, x_3) en coordenadas paralelas (arriba) y cartesianas (abajo).

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 anteriormente en el artículo. Se estudia el rendimiento (medido en 100 kg/ha.) de seis cultivos herbáceos (variables), para quince países (observaciones) que forman parte de la Unión Europea para el año 2002 (*Tabla.1*).

Por medio de la *Figura 8* de Coords ‖, se observa una alta correlación positiva entre la producción de trigo y cebada. Cambiando el orden los cultivos, se revela una correlación moderada cebada-colza y trigo-colza. Bélgica destaca en la producción de maíz, al contrario de Irlanda, Finlandia, Luxembuergo, Dinamarca, Reino Unido y Suecia que no producen este cereal. España es uno de los mayores productores de maíz pero su producción es mucho menor en todos los otros cereales.

El orden de las Coords ‖ es una condición que puede afectar significativamente la expresividad del gráfico, variando el orden de los cereales es posible abreviar el problema sin la reducción del contenido o de la modificación de los datos de alguna manera. Esta organización puede tener un impacto importante en la expresividad de las coordenadas. Los posibles ordenamientos de las dimensiones pueden revelar los aspectos distintos de los datos, obteniéndose conclusiones totalmente diferentes.

Tabla 1. Rendimiento de cultivos herbáceos de los países de la Unión Europea, 2002 (100 kg./ha).

Países	Trigo	Cebada	Centeno	Maíz	Colza	Girasol
Alemania	75.4	60.2	57.8	87.8	32.9	25.0
Austria	53.6	47.3	39.3	95.7	28.3	26.2
Bélgica	84.4	73.2	40.0	122.7	35.0	0.0
Dinamarca	70.1	50.5	48.6	0.0	29.0	0.0
España	25.9	23.9	17.1	94.7	13.3	6.8
Finlandia	21.5	27.0	20.0	0.0	14.3	0.0
Francia	72.8	62.2	45.4	89.0	32.7	22.9
Grecia	23.3	26.9	22.1	85.7	0.0	0.0
Irlanda	87.8	66.6	0.0	0.0	33.0	0.0
Italia	34.9	37.7	30.0	97.5	34.0	30.8
Luxemburgo	57.5	52.3	40.0	0.0	27.6	0.0
Países Bajos	83.4	62.8	46.7	70.0	33.3	0.0
Portugal	16.4	13.0	10.2	53.4	0.0	8.0
Reino Unido	80.5	55.8	53.8	0.0	32.3	0.0
Suecia	60.3	38.4	46.8	0.0	21.1	0.0

Fuente: Agriculture in the European Union: Statistical and Economic information, 2003.

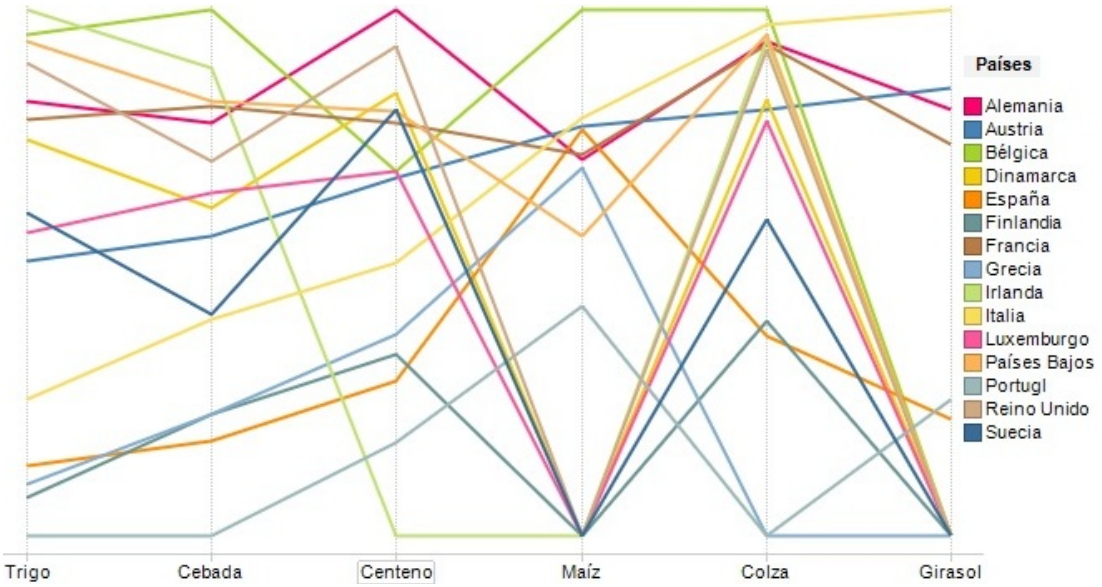

Figura 8. Representación en Coords ‖ del rendimiento de cultivos herbáceos en los países de la Unión Europea, 2002.

CONCLUSIONES

Hemos mostrado que las Coords ‖ son un método novedoso para visualizar un plano n-dimensional en una representación de dos dimensiones y que son especialmente útiles 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 observar 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.

Por otra parte, es una herramienta que nos ayuda a detectar los valores extremos multivariantes presentes en conjuntos 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.

European Commission (2003). Agriculture in the European Union: Statistical and Economic information. Office for Official Publications of the European Communities, Luxembourg.

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.; 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.