

Un dispositivo gráfico Biplot para modelar relaciones entre dos conjuntos de variables

A Biplot graphical tool to model the relationships between two sets of variables

Maura Vásquez*, **Guillermo Ramírez***,
Alberto Camardiel* y **Tomás Aluja****

Recibido: 08-01-08 / Revisado: 16-02-08 / Aceptado: 25-03-08

Códigos JEL: C1, C10

Resumen

En la regresión multivariante se analizan relaciones entre un conjunto de indicadores de un determinado fenómeno (variables y 's) y otro de variables explicativas (variables x 's). Previa a la modelación resulta útil contar con información clave sobre las relaciones entre ambos. En este sentido, se presenta un gráfico desarrollado por Ter Braak (1990) y por Ramírez *et al.* (2005) en forma independiente. Se proponen indicadores de la calidad de las aproximaciones: a) variables y 's bien explicadas por las x 's, b) agrupaciones de variables y 's cuyas estimaciones presentan fuertes correlaciones y c) subconjuntos de variables explicativas que presentan mayor correlación con una variable y particular. El gráfico presentado es un GH-Biplot generalizado (Gabriel, 1971) que utiliza variables latentes generadas por un análisis de correlación canónico. Se evalúan sus alcances en una aplicación en la que se explica el comportamiento de los indicadores que conforman el Índice de Desarrollo Humano en términos de otras variables del entorno social y económico a nivel estatal en Venezuela.

Palabras clave: Biplot, análisis canónico, regresión multivariante, desarrollo humano.

Abstract

Multivariate multiple regression analysis is used to model the relationships between two sets of variables, the y -set of response variables and the x -set of explaining variables. In trying to have useful information about the nature of the relationships between the two sets, Ter Braak (1990) proposed a graphical display taken up again by Ramírez *et al.* (2005) in an independent way. In this paper, indexes of the quality approach are considered: a) those y variables better explained by x variables, b) clusters of y variables whose predictions are

* Postgrado en Estadística, Universidad Central de Venezuela (UCV), e-mails: mauravasquez@cantv.net, guiram@cantv.net, acamar@reacciun.ve

** Universidad Politécnica de Cataluña, España, e-mail: tomas.aluja@upc.edu

strongly correlated and c) subsets of x variables more correlated with a specific y variable. The graphical device considered is a generalized GH-Biplot (Gabriel, 1971) using latent variables generated by a canonical correlation analysis. The graphical display was used to analyze the Human Development Index corresponding to the 23 states of Venezuela related to several social and economic indicators.

Key words: Biplot, canonical analysis, multivariate regression, human development.

1. Introducción

El análisis de regresión múltiple multivariante (ARMM) es aplicable a un sinnúmero de fenómenos de la realidad en los que se dispone de dos bloques diferentes de variables, entre los cuales se establecen relaciones de carácter asimétrico: un primer bloque de variables denotado por $Y = \{y_1, y_2 \dots y_q\}$ que describe el comportamiento del fenómeno propiamente dicho, y otro de variables explicativas denotado por $X = \{x_1, x_2 \dots x_p\}$. En una fase previa a la modelación resulta de gran utilidad contar con gráficos que den cuenta de aspectos claves de las relaciones entre los dos conjuntos.

En este trabajo se presenta un dispositivo gráfico desarrollado por Ter Braak (1990) y por Ramírez *et al.* (2005) en forma independiente, que ha sido diseñado con el propósito de identificar las variables y_k mejor explicadas por las x 's, las variables y_k estimadas en términos de las x 's, que pudieran considerarse redundantes y subconjuntos de las x 's con mayor correlación con una y_k particular. Se derivan indicadores de la calidad de las aproximaciones obtenidas y se evalúan sus alcances en una aplicación en la que se explica el comportamiento de los indicadores que integran el Índice de Desarrollo Humano (IDH) en función de un conjunto de variables del entorno social y económico en Venezuela a nivel estatal, con la cual se ilustra el procedimiento gráfico en consideración.

2. Fundamentación teórica

Diversos autores como Wold (1983) y Ter Braak (1998) discuten una perspectiva para el ARMM, denominada regresión PLS (*Partial Least Squa-*

res Regression), Regresión con Mínimos Cuadrados Parciales, en la cual se propone encontrar variables latentes para el conjunto X y el conjunto Y simultáneamente, que sean relevantes en la explicación de la estructura de covarianzas entre los dos bloques. En concordancia con este planteamiento, el dispositivo gráfico propuesto en este trabajo ha sido diseñado para destacar los aspectos más importantes del comportamiento de las y 's en términos de sus relaciones con el conjunto de las x 's, utilizando las variables latentes generadas en el análisis de correlación canónico (ACC), las cuales poseen alta potencialidad para describir las intercorrelaciones entre los dos conjuntos de variables.

El dispositivo propuesto es un gráfico tipo Biplot (Gabriel, 1971) basado en la descomposición en valores singulares generalizada (DVSG) de la matriz R_{xy} , arreglo de rango s que contiene información sobre el grado y dirección de las relaciones lineales entre las variables de uno y otro bloque. Además, su factorización DVSG puede escribirse como un agregado de matrices, cada una de las cuales es construida con información reportada por las correlaciones de las x 's y de las y 's con las variables latentes de su grupo respectivo generadas por el ACC (Tenenhaus, 1998):

$$\begin{aligned} R_{xy} &= ND_{\sqrt{\lambda}} M^t = \sum_{h=1}^s \sqrt{\lambda_h} (N_h)(M_h)^t \\ &= \sum_{h=1}^s r_h (R_{xx} a_h)(R_{yy} b_h)^t \end{aligned} \quad (1)$$

donde:

- R_{xx} , R_{yy} son matrices de correlaciones dentro de los bloques X e Y respectivamente.
- a_h y b_h son los vectores de coeficientes del h -ésimo par de variables canónicas $t_h = Xa_h$ y $u_h = Yb_h$, y r_h es el coeficiente de correlación canónico entre ellas.
- Las coordenadas del vector $N_h = R_{xx} a_h$ son las correlaciones de las variables explicativas con la h -ésima variable canónica de su grupo (cargas canónicas):

$$= (r(x_1, t_h) \dots r(x_j, t_h) \dots r(x_p, t_h)) \quad h = 1, 2, \dots, s \quad (2)$$

Y las coordenadas del vector $M_h = R_{yy} b_h$ se expresan en términos de las correlaciones de las y 's con la h -ésima variable canónica de su respectivo grupo, o con la variable canónica del bloque de las x 's, una vez multiplicadas por r_h (cargas canónicas cruzadas):

$$\begin{aligned} r_h (R_{yy} b_h)^t &= r_h (r(y_1, u_h) \dots r(y_k, u_h) \dots r(y_q, u_h)) \\ &= (r(y_1, t_h) \dots r(y_k, t_h) \dots r(y_q, t_h)) \end{aligned} \quad (3)$$

d) Los autovectores de la factorización DVSG de R_{xy} son ortonormalizados con las métricas R_{xx}^{-1} y R_{yy}^{-1} , respectivamente:

$$(R_{xx} a_\alpha)^t R_{xx}^{-1} (R_{xx} a_\beta) = (R_{yy} b_\alpha)^t R_{yy}^{-1} (R_{yy} b_\beta) = \begin{cases} 1 & \text{si } \alpha = \beta \\ 0 & \text{si } \alpha \neq \beta \end{cases}$$

e) La descomposición DVSG de R_{xy} se obtiene a partir de la DVS de $R_{xx}^{-1/2} R_{xy} R_{yy}^{-1/2}$ (Greenacre, 1984):

$$R_{xx}^{-1/2} R_{xy} R_{yy}^{-1/2} = L D_{\sqrt{\lambda}} V^t$$

siendo $L = R_{xx}^{-1/2} N$ y $V = R_{yy}^{-1/2} M$, tales que $L^t L = V^t V = I$.

En consecuencia, las coordenadas de la DVSG de R_{xy} quedan en la forma:

- Coordenadas de las filas: $N = R_{xx}^{1/2} L$ y
- Coordenadas de las columnas: $M D_{\sqrt{\lambda}}^{-1/2} = R_{yy} V D$

3. Propuesta

En este trabajo se evalúan algunas propiedades del dispositivo gráfico cuyo propósito fundamental consiste en representar de manera aproximada la información contenida en la matriz R_{xy} , lo que permite visualizar la estructura de las relaciones de dependencia lineal de las y 's en términos de las x 's. Específicamente en esta representación pueden identificarse:

- La proporción de variabilidad de cada y_k que es explicada por el conjunto de las x 's.
- Fuerza y dirección de las relaciones lineales entre las estimaciones de y_k e y_l obtenidas en función de las x 's.
- Subconjuntos de variables x 's más correlacionadas con una y_k particular.

La representación biplot es definida por la factorización GH-biplot generalizada (Gabriel, 1971) de R_{xy} :

$$R_{xy} = G_{pxr} (H^t)_{rxq}$$

donde $G = N(s)$ es la matriz de marcadores para las filas y $H = M(s)D_{\sqrt{\lambda}}(s)$ es la matriz de marcadores para las columnas, con las restricciones:

$$(N(s))^t R_{xx}^{-1} (N(s)) = (M(s))^t R_{yy}^{-1} (M(s)) = I$$

De esta manera, los marcadores g_j utilizados para la representación de las filas (las x 's) quedan definidos por las cargas canónicas de las x 's con las variables latentes de su grupo:

$$g_i^t = (r(x_j, t_1) \dots r(x_j, t_h) \dots r(x_j, t_s)) \quad (4)$$

y los marcadores h_k correspondientes a la representación de las columnas (las y 's) quedan a su vez definidos por sus correlaciones con las variables latentes del espacio de las x 's:

$$\mathbf{h}_k^t = (\mathbf{r}(y_k, t_1) \dots \mathbf{r}(y_k, t_h) \dots \mathbf{r}(y_k, t_s)) \quad (5)$$

La representación gráfica de los marcadores así definidos garantiza la mejor aproximación mínimo-cuadrática de rango s para R_{xy} , ya que:

- 1) El producto interno entre marcadores fila y columna reproduce la correlación entre las variables x_j e y_k :

$$\mathbf{g}_j^t \mathbf{h}_k = \sum_{h=1}^s \mathbf{r}(x_j, t_h) \mathbf{r}(y_k, t_h) = \mathbf{r}(x_j, y_k) \quad (6)$$

Esta reconstrucción evidencia, en forma análoga al análisis de factores, que la correlación entre las variables x_j e y_k es determinada por sus relaciones comunes con las variables latentes del espacio de las x 's.

- 2) La norma cuadrado del marcador columna asociado con y_k , cuando se consideran las s variables latentes existentes, representa la proporción de su variabilidad que es explicada conjuntamente por las x 's. En el caso en que se trabaje con una aproximación de rango $r < s$, se aproxima el índice de redundancia de Stewart y Lowe, que se interpreta como la proporción de variabilidad de y_k explicada por las r primeras variables latentes:

$$\|\mathbf{h}_k\|^2 = \sum_{h=1}^s \mathbf{r}^2(y_k, t_h) = R^2 y_k \cdot t_1 t_2 \dots t_r \quad (7)$$

De este manera, marcadores h_k con longitudes grandes corresponden a variables y_k cuya variabilidad queda bien determinada por las variables explicativas, mientras que aquéllas de corta longitud constituyen un indicativo de variables no bien explicadas por las x 's y por tanto podrían ser eventualmente eliminadas.

- 3) El coseno del ángulo entre los marcadores columna h_k y h_l representa una medida que describe la correlación entre los vectores de estimaciones de y_k y y_l obtenidos en la regresión con las x 's.

$$\cos(\text{angulo}(h_k, h_1)) = \frac{Ry_k Y_1 \cdot \{x_1 \dots x_p\}}{Ry_k \cdot \{x_1 \dots x_p\} Ry_1 \cdot \{x_1 \dots x_p\}} \quad (8)$$

- 4) La proyección de g_j sobre h_k reproduce un ordenamiento de las x 's en términos de la magnitud de su correlación con y_h .

En efecto como la proyección de g_j sobre h_k es de la forma:

$$\text{proy}(g_j / h_k) = \frac{g_j^t h_k}{h_k^t h_k} h_k = \frac{r(x_j, y_k)}{R^2 y_k \cdot x_1 x_2 \dots x_p} h_k \quad \text{para } j = 1, 2, \dots, p$$

resulta claro que el ordenamiento de estas proyecciones depende de la magnitud y dirección de la relación lineal entre las x_j y la variable y_k .

- 5) La calidad de reconstrucción de la medida de redundancia* sobre el primer plano biplot, queda medida por el cociente:

$$\frac{\text{Rd}(Y, t_1, t_2)}{\text{Rd}(Y, \{t_1 \dots t_s\})} = \frac{\sum_{h=1}^2 \sum_{k=1}^q r^2(y_k, t_h)}{\sum_{h=1}^s \sum_{k=1}^q r^2(y_k, t_h)}$$

$$* \text{Rd}(Y, t_h) = \sum_{k=1}^q r^2(y_k, t_h) / q$$

(Proporción promedio de variabilidad de las Y 's explicada por t_h)

(Si rango (R_{xy}) = p, entonces

$$R^2 y_k \cdot x_1 x_2 \dots x_p = \sum_{h=1}^s \sum_{k=1}^q r^2(y_k, t_h)$$

4. Información captada por el primer plano R_{xy} BIPLLOT

La información captada por el primer plano R_{xy} biplot queda fundamentada en los siguientes lemas generales, en algunos de los cuales se asumen resultados presentados en (Tenenhaus, 1998):

Lema 1. Medida global de la calidad de las aproximaciones.

El cociente:

$$\frac{\sum_{\alpha=1}^2 \lambda_{\alpha}}{\sum_{\alpha=1}^s \lambda_{\alpha}} = \frac{\sum_{h=1}^2 \Gamma_h^2}{\sum_{h=1}^s \Gamma_h^2} \tag{9}$$

es un indicador de la bondad de las aproximaciones R_{xy} biplot y puede interpretarse como una medida de la proporción de varianza compartida por los conjuntos X e Y que es captada por el primer plano biplot.

Demostración

La aproximación mínimo cuadrática generalizada de rango 2 para R_{xy} :

$$R_{xy}(2) = N(2)D_{\sqrt{\lambda}}(2)M^t(2)$$

minimiza la expresión:

$$\begin{aligned} & \text{tr}(R_{xx}^{-1}(R_{xy} - R_{xy}(2))R_{yy}^{-1}(R_{xy} - R_{xy}(2))^t) = \\ & \text{tr}(R_{xx}^{-1}(ND_{\sqrt{\lambda}}M^t - N(2)D_{\sqrt{\lambda}}(2)M^t(2))R_{yy}^{-1}(ND_{\sqrt{\lambda}}M^t - N(2)D_{\sqrt{\lambda}}(2)M^t(2))^t) = \\ & \text{tr}((LD_{\sqrt{\lambda}}V^t - L(2)D_{\sqrt{\lambda}}(2)V^t(2))(LD_{\sqrt{\lambda}}V^t - L(2)D_{\sqrt{\lambda}}(2)V^t(2))^t) = \\ & \text{tr}((L(s-2)D_{\sqrt{\lambda}}(s-2)V^t(s-2))(L(s-2)D_{\sqrt{\lambda}}(s-2)V^t(s-2))^t) = \\ & \text{tr}((V(s-2)D_{\sqrt{\lambda}}(s-2)L^t(s-2))(L(s-2)D_{\sqrt{\lambda}}(s-2)V^t(s-2))^t) = \\ & \text{tr}(V(s-2)D_{\lambda}(s-2)V^t(s-2)) = \\ & \text{tr}(V^t(s-2)V(s-2)D_{\lambda}(s-2)) = \end{aligned}$$

$$\sum_{\alpha=3}^s \lambda_{\alpha}$$

que es una medida del desfase de la aproximación.

Lema 2. Calidad de las aproximaciones en el espacio de las y 's

El cociente:

$$cr(t_1, t_2) = \frac{\sum_{\alpha=1}^2 r^2(y_k, t_h)}{R^2_{y_k \cdot x_1 x_2 \dots x_p}} = \frac{\sum_{\alpha=1}^2 h_{k\alpha}^2}{\sum_{\alpha=1}^s h_{k\alpha}^2} \quad (10)$$

es la proporción de la variabilidad de y_k explicada por las x 's que es captada por el primer plano biplot, definido por las dos primeras variables latentes del ACC.

Demostración

$$HH^t = (MD_{\sqrt{\lambda}})(MD_{\sqrt{\lambda}})^t = MD_{\sqrt{\lambda}} N^t R_{xx}^{-1} N D_{\sqrt{\lambda}} M^t = R_{yx} R_{xx}^{-1} R_{xy}$$

que es una matriz cuyos elementos sobre la diagonal principal son $R^2_{y_k \cdot x_1 x_2 \dots x_p}$ y en la posición (k, l) $(k, l) R_{y_k y_l \cdot x_1 x_2 \dots x_p}$.

Por consiguiente:

$$\sum_{\alpha=1}^s h_{k\alpha}^2 = \sum_{\alpha=1}^s r^2(y_k, t_\alpha) = y_k^t X(X^t X)^{-1} X^t y_k$$

5. Aplicación

El dispositivo gráfico ha sido utilizado para analizar el comportamiento de los indicadores de Salud (Esperanza de vida (EDV)), Educación (Alfabetismo (TALF) y Tasa de matrícula combinada (MCOM)) e Ingreso (Ingreso per cápita en dólares PPA), correspondientes a las 23 entidades federales en Venezuela durante el año 2001. La información ha sido tomada del *Atlas de Desarrollo Humano en Venezuela* publicado por el Instituto Nacional de Estadística. Los cálculos y gráficos han sido obtenidos.

nidos mediante un programa en lenguaje MatLab, versión 2006, elaborado por los autores.

Este conjunto de indicadores –variables Y – con los cuales se obtiene el IDH a nivel estatal será descrito en términos de sus relaciones con el siguiente conjunto de variables X : % de población en la entidad (%POB), densidad de población (DENS), razón de dependencia económica (RDEF), tasa de fertilidad (FERT), tasa de matrícula en preescolar (MPRE), tasa de matrícula en básica (MBAS), tasa de matrícula en media (MMED), tasa de matrícula en diversificada (MDIV), tasa de deserción en básica (DBAS), tasa de deserción en media (DMED), tasa de mortalidad neonatal (MNEO), tasa de mortalidad infantil (MINF), habitantes por hospital (HHOS), inversión en educación (IEDU), inversión en salud (ISAL), delitos contra las personas (DPER) y delitos contra la propiedad (DPRO). En el cuadro 1 se puede observar que un 56,4% de la variabilidad compartida por las y 's y las x 's queda reflejada sobre el primer plano biplot, y se incrementa a un 83% si se consideran las tres primeras direcciones.

Cuadro 1. Variabilidad compartida explicada

Factor	λ_{α}	%	% cum
1	0,9980	28,5	28,5
2	0,9791	27,9	56,4
3	0,9178	26,2	82,6
4	0,6113	17,4	100,0
	3,5062	100,0	

Del cuadro 2 se desprende que la variabilidad de las y 's explicada por las x 's es importante. En términos globales, se tiene que el conjunto de las x 's, a través de las dos primeras variables latentes, explica un 78% de la variabilidad conjunta de los indicadores de desarrollo humano, mientras que al considerar cada variable en forma individual, se observan proporciones de variabilidad explicada que oscilan entre 0,57 y 0,95 (columna III). Los coeficientes de calidad de representación (columna V) indican

que la variabilidad de MCOM, EDV y \$PPA explicada por las x 's queda representada en más del 79% sobre el plano.

Cuadro 2. Índices de redundancia

Variable	II Factor		III Redundancia	IV Coeficiente de determinación	V Calidad de representación
	1	2	$R^2_{y_k.\{t_1,t_2\}}$	$R^2_{y_k.\{x_1\dots x_p\}}$	cr(2)
TALF	0,7174**	0,2416	0,5730	0,9358	0,61
MCOM	-0,4900*	0,7113**	0,7460	0,9492	0,79
EDV	0,6182**	0,7545**	0,9514	0,9772	0,97
\$PPA	0,7822**	0,4912*	0,8531	0,9581	0,89
red(Y, α)	0,4372	0,3437	0,7809		

* significativo al 5%

** significativo al 1%

De los cuadros 2 y 3 se desprende que la primera variable latente puede considerarse como un indicador de calidad de vida que contrasta, por un lado, elevados índices de ingreso, alfabetismo y salud, asociados con alta concentración de población, bajos niveles de dependencia económica y problemas de inseguridad; y por el otro, valores de estos índices que reflejan un desarrollo humano precario.

Cuadro 3. Correlaciones de las x 's con las variables latentes

Variable	1	2	Variable	1	2
%POB	0,5280**	0,3597	DMED	-0,4201*	-0,4504*
DENS	0,1851	0,4726*	MNEO	-0,0673	0,0642
TDEP	-0,6121**	-0,6362**	MINF	-0,5362**	-0,7893**
FERT	-0,7415**	-0,3957	HHOS	0,2317	-0,4828
MPRE	0,0245	0,0453	IEDU	-0,1277	-0,1132
MBAS	-0,7037**	0,3687	ISAL	0,3303	0,3207
MMED	-0,0802	0,6505**	DPER	0,2876	0,3265
DBAS	-0,2389	-0,1412	DPRO	0,6088**	0,5298**

* significativo al 5%

** significativo al 1%

La figura 1 resume la estructura de correlaciones entre los valores estimados de las y 's, en la que se destaca la existencia de dos grupos de variables: (esperanza de vida, ingreso y alfabetismo) con correlaciones positivas importantes y (matrícula combinada) incorrelacionado con el anterior. Se reflejan además en el gráfico las asociaciones entre estos dos grupos y las variables explicativas. Vale la pena destacar que los indicadores de entorno social considerados no tienen capacidad explicativa sobre el comportamiento de la matrícula combinada, lo que se deduce de la magnitud de sus proyecciones sobre el vector que representa a la matrícula combinada.

Los resultados indican que las variables explicativas (porcentaje de población, tasa de mortalidad infantil, tasa de fertilidad, razón de dependencia económica y delitos contra la propiedad), la primera en relación directa y las otras tres en relación inversa, resultan suficientes para predecir salud, ingreso y educación, mientras que (tasa de matrícula en básica y tasa de deserción en media) son las variables x 's más importantes para explicar la tasa de matrícula combinada.

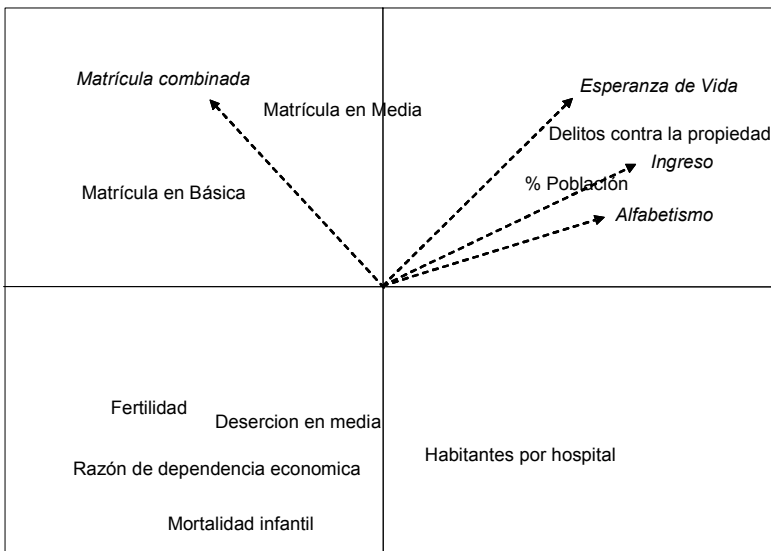


Figura 1. Primer plano factorial R_{xy} biplot

Los resultados indican que las variables explicativas (porcentaje de población, tasa de mortalidad infantil, tasa de fertilidad, razón de dependencia económica y delitos contra la propiedad), la primera en relación directa y las otras tres en relación inversa, resultan suficientes para predecir salud, ingreso y educación, mientras que (tasa de matrícula en básica y tasa de deserción en media) son las variables x 's más importantes para explicar la tasa de matrícula combinada.

6. Conclusiones

Por construcción, el dispositivo presentado aquí permite cumplir con el propósito de lograr representaciones gráficas aproximadas de algunos aspectos claves de las relaciones entre dos conjuntos de variables, y condensar simultáneamente en un solo gráfico información que, de otra manera, debería ser analizada mediante sucesivas tablas, contentivas de la misma información, pero de manera desconectada.

A partir de la representación en algún plano biplot es posible disponer de información aproximada para evaluar visualmente la proporción de variabilidad de cada una de las variables y 's que es explicada conjuntamente por las variables predictoras, a través de las variables latentes cuyas direcciones definen el plano. Este resultado no solamente proporciona una medida esencial de la fuerza con la que las variables x 's explican el comportamiento de cada una de las variables en el otro conjunto, sino que, además, es posible entender los mecanismos subyacentes que orientan las direcciones principales de la estructura de relaciones entre las x 's para explicar el comportamiento de las y 's. Es importante destacar además que el dispositivo ha sido dotado complementariamente de un conjunto de medidas que dan cuenta de la calidad de la información sobre él representada.

7. Referencias

- Gabriel, Karl Ruben (1971). "The Biplot graphic display of matrices with application to principal component analysis." *Biometrika*, 58, pp. 453-467.
- Greenacre, Michael. (1984): *Theory and Applications of Correspondence Analysis*, London: Academic Press.
- Instituto Nacional de Estadística (2004). *Atlas del Desarrollo Humano en Venezuela*. Caracas: Publicación Electrónica.
- Ramírez, Guillermo, Maura Vásquez y Alberto Camardiel (2005). *A graphical tool Biplot based on the SVD of R_{xy} to enhance PLS model construction*. Proceedings of the 4th International Symposium on PLS and Related Methods, Barcelona.
- Tenenhaus, Michel. (1998). *La Régression PLS: Théorie et Pratique*, Paris: Ed. Technip.
- Ter Braak, Cajo (1990). "Interpreting canonical correlation analysis through biplots of structure correlations and weights." *Psychometrika*, 55, 3, pp. 519-531.
- Wold Svante, Martens Harald y Wold Herman (1983). *The multivariate calibration method in chemistry solved by the PLS method*. Heidelberg: Ruhe Kagstrom Eds., Proc. Conf. Matrix Pencils, Lecture Notes in Mathematics, Springer-Verlag, pp. 286-293.