Psicothema 2007. Vol. 19, nº 3, pp. 497-505
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Desarrollo de técnicas de visualización múltiple en el programa ViSta: ejemplo de aplicación al análisis de componentes principales

Rubén Ledesma, J. Gabriel Molina*, Forrest W. Young (†)** y Pedro Valero-Mora*
CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas) (Argentina), * Universidad de Valencia y ** University of North Carolina at Chapel Hill (USA)

La visualización múltiple (VM) es una técnica gráfica de análisis de datos que cuenta con escasa difusión en la práctica a pesar de su potencial aplicado aparente. En este trabajo: (1) se describe la VM como técnica gráfica aplicada al contexto del análisis estadístico de datos; (2) se plantean una serie de principios relativos al diseño de una VM; (3) se muestra el esquema general de desarrollo de una VM en un entorno informático concreto, el sistema estadístico ViSta; (4) se ilustra este desarrollo a través de un ejemplo de VM aplicada al análisis de componentes principales; y, por último, (5) se discuten algunas cuestiones asociadas al desarrollo y aplicación de la VM como técnica gráfica.

Multiple visualisation in data analysis: A ViSta application for principal component analysis. Multiple visualisation (MV) is a statistic graphical method barely applied in data analysis practice, even though it provides interesting features for this purpose. This paper: (1) describes the application of the MV graphical method; (2) presents a number of rules related to the design of an MV; (3) introduces a general outline for developing MVs and shows how MV may be implemented in the ViSta statistical system; (4) illustrates this strategy by means of an example of MV oriented to principal component analysis; and, finally, (5) discusses some limitations of using and developing MVs.

Los programas informáticos orientados al análisis de datos representan uno de los desarrollos más relevantes producidos en el campo del software para fines científicos. En Psicología, programas como *SPSS* (Spss Inc., 2006) o *Statistica* (StatSoft Inc., 2005), entre otros, constituyen hoy en día herramientas básicas de trabajo en la investigación. Estos sistemas informáticos han incrementado la potencia y flexibilidad del análisis estadístico y han puesto a disposición del investigador la posibilidad de aplicar un repertorio amplio de técnicas de diferente cometido y complejidad, la mayoría de las cuales pueden ser ejecutadas haciendo uso de interfaces gráficas simples e intuitivas.

Por otro lado, el desarrollo informático también ha permitido la creación de técnicas gráficas cada vez más sofisticadas, dinámicas e interactivas para visualizar datos y/o resultados de modelos estadísticos. Estas técnicas, si bien se han visto impulsadas desde el análisis exploratorio de datos (Tukey, 1977, 1980), no acaban de incorporarse al software de uso más habitual, por lo menos en lo referente a los desarrollos más sofisticados, como pueden ser los gráficos *dinámicos* e *interactivos* (Young, Valero-Mora, y Friendly, 2006). Esto explicaría, en parte, porqué ciertas técnicas gráficas modernas, a pesar de su potencial interés, son poco conocidas y utilizadas por los investigadores. De hecho, estos desarrollos sólo se encuentran implementados en programas menos difundidos y utilizados en Psicología, como DataDesk (DataDescription Inc., 2005), GGobi (Swayne, Lang, Buja, y Cook, 2003) o ViSta (Young, 2006).

Este trabajo trata sobre un tipo de técnica grafica que, a pesar de su interés y utilidad potencial, se encuentra escasamente difundida en Psicología: la *Visualización Múltiple* (VM). El trabajo se organiza del siguiente modo: primero se presenta una revisión sobre los llamados *gráficos dinámicos* e *interactivos*; a continuación se ofrece una descripción más en detalle de la técnica de la VM; por último, se describe cómo ésta puede implementarse en el sistema ViSta, a la vez que se ilustra dicha implementación mediante un ejemplo aplicado al análisis de componentes principales (ACP), un tipo de método ampliamente utilizado en Psicología (Rivas y Martínez-Arias, 1991). Se pretende, en definitiva, contribuir a una mayor difusión y utilización de la técnica gráfica de la VM en el análisis de datos, a la vez que alentar la implementación de la misma entre los expertos en análisis de datos y programación estadística.

Las técnicas gráficas dinámicas e interactivas

Varios autores coinciden en afirmar que los denominados *gráficos dinámicos* e *interactivos* constituyen el último paso importante en el desarrollo de la estadística gráfica (véase, por ejemplo, Friendly, 2006; Wainer y Velleman, 2001). A diferencia de los gráficos estáticos convencionales, utilizados principalmente con fines de presentación, comunicación o simplemente decorativos, estas técnicas permiten que el analista pueda interactuar con las representaciones gráficas y, de ese modo, realizar una exploración más activa de sus datos (Wainer y Velleman, 2001).

Fecha recepción: 1-6-2006 • Fecha aceptación: 16-1-2007
Correspondencia: J. Gabriel Molina
Facultad de Psicología
Universidad de Valencia
46010 Valencia (Spain)
E-mail: gabriel.molina@uv.es

Los gráficos dinámicos e interactivos han encontrado su fundamento e inspiración en la filosofía del análisis exploratorio de datos (Tukey, 1977, 1980), siendo el propio Tukey quien dio uno de los primeros pasos en el desarrollo de estas técnicas: su programa *Prim-9* (Fisherkeller, Friedman, y Tukey, 1975) representa el primer sistema interactivo de visualización en tres dimensiones y un prototipo para los futuros desarrollos en el área. *Prim-9* se diferenciaba claramente de los gráficos precedentes tanto por sus características dinámicas, principalmente la rotación, como por las posibilidades de que el analista interactuara con la representación e incluso controlara aspectos de su dinámica (sentido y orientación de la rotación, etc.) (Friedman y Stuetzle, 2002). Este método supuso una innovación importante en cuanto a la manera de entender el lugar de los gráficos en el análisis de datos, pudiendo considerarse como el primer paso en el devenir de los gráficos 'estáticos' a los gráficos dinámicos e interactivos.

En este sentido, también hay coincidencia en señalar la importancia sustantiva de los trabajos que Cleveland realizara en colaboración con otros autores en materia de sistematización técnica y teórica de los gráficos dinámicos e interactivos (Becker y Cleveland, 1987; Cleveland, 1988; Cleveland y McGill, 1988). Estos trabajos integran varios conceptos y técnicas a partir de los cuales los gráficos pasarían a ser más dinámicos e interactivos para el usuario, mejorando sensiblemente su capacidad exploratoria. Propiedades como la rotación, la identificación de casos en los gráficos («labeling»), el barrido o selección de grupos de casos («brushing»), el ligado empírico de varias representaciones gráficas («linking»), etc., se implementarían como una forma de transformar las imágenes estáticas convencionales en representaciones verdaderamente dinámicas e interactivas (Cleveland y McGill, 1988). La figura 1 ilustra algunos de estos conceptos.

La imagen muestra un ejemplo básico de este tipo de gráficos en funcionamiento, los cuales pueden cambiar o transformarse para brindar diferentes imágenes o visiones de los mismos datos. Tales cambios se producen como respuesta a una acción del usuario sobre el propio gráfico, de modo que éste puede interactuar directamente con la representación gráfica. En el ejemplo, las celdas de la matriz de diagramas de dispersión (izquierda) muestran las relaciones bivariadas entre tres variables, mientras que el *diagrama de dispersión móvil* («Spin-plot») muestra las tres variables simultáneamente. Se puede observar cómo al seleccionar un grupo de observaciones en la matriz, esta acción se propaga al *Spin-plot*, resaltándose en negro las mismas observaciones. Se ilustra así el concepto de selección por barrido (*brushing*) y el concepto de ligado (*linking*), típico de los gráficos dinámicos. Además, se observan en la imagen diferentes opciones (botones *«Zoom», «Speed»,* etc.) para que el usuario interactúe y controle el movimiento del *Spin-plot.*

Tras los trabajos de Cleveland comienzan a desarrollarse, a principios de la década de los noventa, los sistemas informáticos basados en gráficos dinámicos e interactivos, entre cuyos exponentes más importantes encontramos XGobi (actualmente GGobi), DataDesk y ViSta. Estos programas implementan de forma integral conceptos y técnicas previas, pero también extienden las capacidades de los gráficos dinámicos y son utilizados como plataformas para la creación de nuevas y más sofisticadas técnicas gráficas. Por ejemplo, Young, Faldowski y McFarlane (1993) desarrollan en ViSta varias técnicas de VM para propósitos específicos. Esta técnica es heredera de los trabajos previos en materia de gráficos dinámicos e interactivos, pero también supone innovaciones en varios sentidos, sobre todo en lo referente a la integración y coordinación de grupos de gráficos dinámicos.

En lo que respecta al presente, si bien las técnicas gráficas dinámicas e interactivas son la esencia de algunos sistemas estadísticos como XGobi (Swayne et al., 2003) o ViSta (Young, 2006), no ha sucedido lo mismo con algunos programas más extendidos como es el caso de SPSS. Este hecho podría venir motivado por diversas causas: una de ellas, la inercia de los paquetes estadísticos convencionales, probablemente reacios a incorporar técnicas que puedan resultar difíciles de encajar de un modo coherente dentro del cúmulo de procedimientos ya disponibles en el sistema; otra, que estos sistemas suelen estar basados en lenguajes y paradigmas de programación que no soportan gráficos dinámicos e interactivos. Esto último apunta también a ser la causa principal de la estanqueidad de la mayoría de los paquetes estadísticos a la hora de posibilitar la programación de nuevas representaciones gráficas por parte de sus usuarios.

Señalar, por último, que puede encontrarse una revisión detallada del desarrollo histórico de la estadística gráfica en el trabajo de Friendly (2006). Así también, remitimos a Young, Valero-Mo-

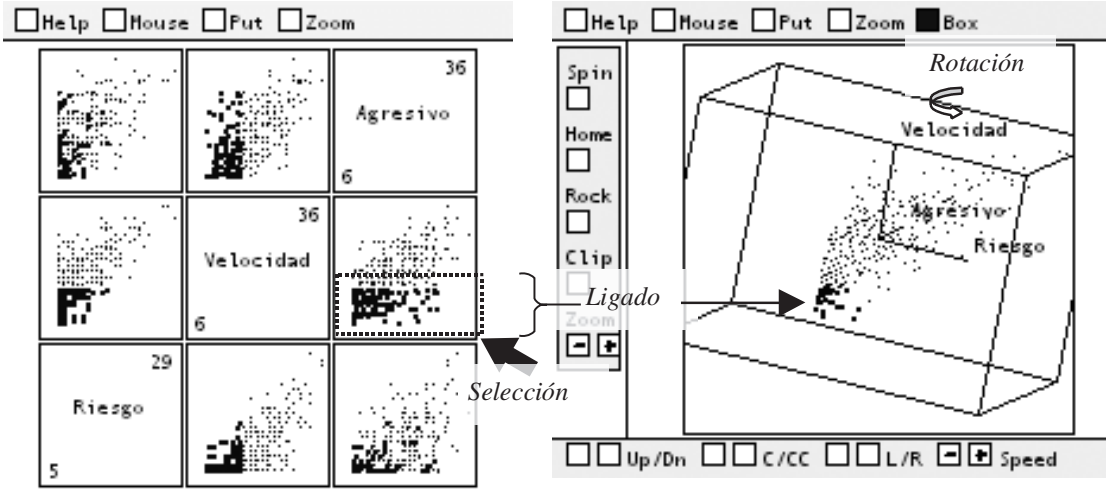

Figura 1. *Ejemplo de gráficos dinámicos e interactivos en funcionamiento. La imagen ilustra tres conceptos básicos: selección, ligado y rotación*

ra y Friendly (2006) para una visión de los desarrollos más recientes en gráficos dinámicos e interactivos.

La técnica gráfica de la VM en el análisis de datos

Conceptualmente, la VM se basa en una idea relativamente sencilla pero potente: se seleccionan una serie de representaciones gráficas de interés en la aplicación de un determinado análisis estadístico; se ligan esas representaciones entre sí de acuerdo a los objetivos de ese modelo de análisis; y se presentan simultáneamente en una misma representación gráfica o visualización, en la que además se integren recursos que permitan al analista interactuar con las mismas (Young et al., 1993). De esta manera, el analista cuenta con diferentes representaciones parciales de la aplicación de un mismo modelo estadístico, pudiendo así aprovechar esa visualización agregada y los vínculos entre las representaciones gráficas para obtener más información de la que podría extraerse considerando esas representaciones de forma independiente. La figura 9, ejemplo en una sección posterior del diseño e implementación de una VM, permite ilustrar ahora la definición de esta técnica.

Una de las aplicaciones potencialmente más útiles de la VM se da en el caso de ciertos modelos o técnicas estadísticas que, por la complejidad o extensión del *output* que generan, tornan difícil la interpretación de resultados a partir de informes numéricos en forma de texto o a partir de representaciones gráficas aisladas. Ello es característico de ciertos modelos de análisis estadístico multivariado, como el análisis factorial, el análisis de conglomerados o el análisis de correspondencias múltiples, por lo que puede resultar especialmente conveniente el desarrollo de VMs que den apoyo a la aplicación de los mismos. En estos casos, crear una VM implicaría, básicamente, *seleccionar, coordinar* y *presentar* un conjunto de gráficos dinámicos e interactivos apropiados para visualizar los resultados del modelo en cuestión.

Aquí, el interés de ViSta radica en que no sólo proporciona al usuario más de 20 VMs para diferentes tipos de datos y modelos estadísticos, sino que también incluye un lenguaje y una arquitectura de programación para la creación de nuevas VMs. Por ello, en este trabajo ViSta ha sido el utilizado para desarrollar una VM concreta que sirviera para ilustrar la aplicación de esta técnica gráfica, a la vez que para poner de manifiesto a los potenciales desarrolladores las posibilidades de este sistema en la creación de VMs.

Recordar que ViSta es un programa gratuito y de código abierto que surgió originalmente para servir como plataforma de desarrollo de nuevos métodos de visualización estadística, si bien representa ya en la actualidad un sistema robusto y con un repertorio amplio de capacidades para gestionar y analizar datos (Molina, Ledesma, Valero-Mora, y Young, 2005). Por último, los interesados pueden encontrar una aproximación técnica a la programación de VMs en ViSta en Young, Valero-Mora, Faldowski y Bann (2003).

Esquema de desarrollo y ejemplo de aplicación de una VM en ViSta

1. Presentación de un ejemplo

En este apartado plantearemos un ejemplo con el fin de ilustrar el desarrollo de una VM en ViSta. Se trata de una aplicación del ACP a datos provenientes de un cuestionario sobre *Estilos de Conducción Vial* denominado MDSI (Ben-Ari, Mikulincer, y Gillath, 2004), aplicado a una muestra de 300 conductores de la ciudad de Mar del Plata, Argentina. La figura 2 muestra una imagen parcial de la tabla de datos del ejemplo en ViSta. Las variables de este ejemplo son las diferentes subescalas del MDSI que evalúan estilos específicos de conducción denominados: Arriesgado; Alta Velocidad; Agresivo; Disociativo; Ansioso; Reducción de Estrés; Seguro; y Cordial. Con estos datos, el ACP es utilizado para explorar y visualizar la estructura de correlaciones entre las diferentes escalas del MDSI.

2. Salida de resultados en formato de listado

Luego de aplicar el ACP a estos datos con ViSta se obtiene una salida en formato texto que se muestra parcialmente en la figura 3. Se trata de una salida similar a la de otros programas estadísticos, incluyendo información sobre el ajuste de la solución, las coordenadas factoriales de las variables y de los casos, etc. En la misma figura también podemos advertir que los resultados del ACP pueden ser extensos y, consiguientemente, difíciles de interpretar de forma directa mediante este tipo de salidas en formato texto. Por ello, algunos sistemas estadísticos ya ofrecen representaciones gráficas que simplifican y ayudan a descifrar parte de la información contenida en estos *outputs*, sea el caso del gráfico de los pesos factoriales de las variables en un espacio bi- o tridimensional de los componentes (gráfico de componentes), o el del gráfico de los autovalores de los componentes (gráfico de sedimentación o «Scree Plot»). La VM desarrollada con ViSta que se presenta en el siguiente apartado muestra cómo se pueden complementar estos

☐Help ☐Save ☐Lock ☐Maximize ☐Expand ☐Format ☐Labels

Type:MulVar	Riesgo	Velocid	Agresio	Disocia	Ansioso	Estres	Prudenc	Cordial
Size:399 X 8	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
Obs1	7.	18.	23.	19.	9.	17.	18.	25.
Obs2	11.	24.	28.	20.	4.	13.	20.	30.
Obs3	15.	31.	21.	15.	6.	20.	16.	25.
Obs4	5.	12.	9.	27.	5.	23.	23.	26.
Obs5	15.	25.	25.	19.	6.	23.	19.	33.
Obs6	17.	25.	25.	28.	11.	18.	21.	27.
Obs7	11.	13.	12.	27.	9.	14.	26.	30.
Obs8	7.	16.	18.	22.	12.	14.	23.	32.

Figura 2. *Imagen parcial de la planilla de datos de ViSta con los datos del ejemplo*

métodos gráficos puntuales con otros y, sobre todo, pretende mostrar cómo éstos se pueden integrar en una visualización que, de un modo global y dinámico, haga factible una *conversación* más interactiva e intuitiva entre el analista y los resultados del ACP.

3. Pasos en el diseño e implementación de una VM

La implementación de una VM en ViSta consta, básicamente, de seis pasos. Éstos se describen a continuación y se ilustran con el desarrollo de una VM para el ejemplo de referencia. El resultado final de este proceso es una VM (denominada '*spreadplot'* en el entorno de ViSta) como la que puede observarse en la figura 9.

Paso 1: Determinar el objeto estadístico a visualizar

Este paso supone definir la información asociada al modelo estadístico en cuestión que se desea mostrar en la VM a implementar. En general, será conveniente considerar información de diferente índole, sea el caso de la solución proporcionada por el modelo estadístico aplicado, el ajuste de dicho modelo, o aquella que ayude a interpretar la solución obtenida, realizar predicciones, etc. En nuestro caso, el interés se centró en el ACP, siendo la información que fue considerada, por una parte, la relativa al ajuste de la solución (varianza explicada e importancia relativa de cada componente) y, por otra parte, la relativa a la solución factorial obtenida, básicamente, las coordenadas de las variables (*Estilos de conducción*) y de los sujetos (*Conductores*) en los componentes extraídos.

Paso 2: Seleccionar un conjunto representativo de gráficos

Se trata de seleccionar un número reducido de gráficos para mostrar la información sobre el resultado del modelo estadístico. Debe procurarse que estas representaciones sean complementarias, a fin de evitar redundancias y aumentar la capacidad informativa de la VM. Así, la selección de estas representaciones gráficas debe favorecer a aquellas que se considere más eficientes en cuanto a su capacidad para transmitir información de un modo intuitivo e interactivo. Por otro lado, pueden elegirse representaciones gráficas innovadoras pero, por supuesto, también aquellas que han sido tradicionalmente utilizadas en la realización de un análisis y que han puesto de manifiesto su eficacia en tal cometido. No obstante, será el desarrollador de la VM quien, en último término, sea el responsable de darle la impronta que le caracterice como método de análisis.

En nuestro caso, las siguientes cinco representaciones gráficas fueron seleccionadas como integrantes de la VM para el modelo de ACP:

(1) Gráfico de sedimentación o *scree-plot* (Cattell, 1966), representación gráfica orientada a evaluar la importancia relativa de cada uno de los componentes principales extraídos, tradicionalmente utilizada a la hora de decidir el número de factores a retener. En nuestro caso, se implementó esta representación gráfica añadiendo a la misma una línea complementaria que proporciona un criterio adicional para decidir el número de factores a conservar (fi-

```
Principal Components Analysis of Variable Correlation

Model:     PrnCmp
Variables: (Riesgo Velocidad Agresivo Disociativo Ansioso Estres Seguro Cordial)

Fit Measures for each Component:
Eigenvalue (amount of total data variance fit by each component)
Proportion (of total data variance fit by each component)
Cumulative Proportion (of total data variance fit by the components)

                FIT MEASURES
COMPONENTS      E-Value      Prop.     CumProp
PC1             3.34277    0.41785    0.41785
PC2             1.29490    0.16186    0.57971
PC3             1.04142    0.13018    0.70989
PC4             0.68838    0.08605    0.79593
PC5             0.50341    0.06293    0.85886
PC6             0.47373    0.05922    0.91808
PC7             0.37450    0.04681    0.96489
PC8             0.28089    0.03511    1.00000

Coefficients (Eigenvectors):
                COMPONENTS
VARIABLES          PC1        PC2        PC3        PC4        PC5        PC6
Riesgo          0.4220    -0.2149    -0.2804     0.1909     0.3564     0.2654
Velocidad       0.4310    -0.2385    -0.2400     0.2928     0.0622     0.2595
Agresivo        0.4157    -0.1965    -0.0404     0.2696    -0.6060    -0.3562
Disociativo     0.3059     0.5374     0.1422    -0.1711    -0.4803     0.5460
Ansioso         0.1953     0.6669     0.0610     0.4815     0.2726    -0.4324
Estres          0.2427     0.2274    -0.6445    -0.6120     0.0667    -0.2886
Seguro         -0.3564     0.2651    -0.4535     0.3276     0.0408     0.3874
Cordial        -0.3819    -0.0520    -0.4660     0.2493    -0.4367    -0.1303

Component Scores:
(Left Singular Vectors times Square Root of Eigenvalues)
                COMPONENTS
OBSERVATIONS       PC1        PC2        PC3        PC4        PC5        PC6
Obs1            0.0496    -0.0177     0.0552    -0.0186     0.0135    -0.0561
Obs2            0.0454    -0.0969     0.0067     0.0271    -0.0475     0.0006
Obs3            0.1235    -0.1007    -0.0067    -0.0183     0.0599    -0.0142
Obs4           -0.0326     0.0360     0.0087    -0.1137     0.0257     0.0210
Obs5            0.0721    -0.0709    -0.0821    -0.0057    -0.0208    -0.0328
Obs6            0.1255     0.0044    -0.0146     0.0395     0.0148     0.0072
Obs7           -0.0386     0.0423     0.0043     0.0145     0.0211     0.0381
Obs8           -0.0261     0.0375     0.0095     0.0476    -0.0061    -0.0368
Obs9            0.0129    -0.0295     0.0911    -0.0010    -0.0053    -0.0532
Obs10           0.1615     0.0071    -0.0108    -0.0304     0.0068     0.1071
```

Figura 3. *Imagen parcial de la salida en formato texto de ViSta para el modelo de ACP*

gura 4). Esta línea está basada en el resultado de una función propuesta por Keeling (2000) que aproxima una solución al análisis paralelo clásico (Horn, 1965), evitando el proceso de simulación subyacente al enfoque tradicional. El análisis paralelo es una técnica de simulación que permite obtener autovalores críticos para los componentes principales, con lo cual puede decidirse el número de factores a conservar mediante un criterio más objetivo que otras reglas de uso extendido, como la regla *K1* (Ruiz y San Martín, 1992). En este caso, el criterio gráfico es muy simple: conservar los factores que se ubican por encima de la línea de autovalores nulos estimados. Así, para los datos del ejemplo, el gráfico sugiere la extracción de dos componentes principales para la interpretación, los cuales explicarían el 57.97% de la varianza total.

(2) Diagrama de dispersión con las variables y los casos representados simultáneamente en el espacio de los componentes principales. En este gráfico, conocido en la literatura como *Bi-plot* (Gabriel, 1981), se representan las variables como vectores y los individuos (conductores) como puntos en un espacio bidimensional de los componentes principales. En el *Bi-plot* incluido en nuestra VM (figura 5), el analista puede determinar en todo momento, y sin abandonar la VM, los componentes que van a definir los ejes de abscisas y ordenadas de la representación gráfica, aunque aparecerán representados por defecto los dos primeros componentes principales extraídos. En este caso, podemos visualizar la manera en que se agrupan/oponen los estilos de conducción —variables— en la solución. Así, el primer eje agrupa los estilos *Riesgo*, *Velocidad* y *Agresión* (coordenadas positivas a la derecha) y los opone a los estilos *Prudente* y *Cordial* (coordenadas negativas a la izquierda). Por su lado, el segundo eje (orientación vertical) agrupa los estilos *Ansioso* y *Disociativo* de las escalas del MDSI.

(3) Versión tridimensional del *Bi-plot*, un diagrama equivalente al anterior pero tridimensional y dinámico (figura 6). En éste el analista va a contar con una serie de herramientas (barras de botones vertical y horizontal inferior) que le permitan alcanzar diferentes perspectivas de la nube de puntos (conductores) y vectores (variables) representados, ello de forma inmediata a su manipulación. La visualización de la nube de puntos desde diferentes perspectivas ofrece al usuario la posibilidad de aproximarse a una imagen tridimensional de la misma y, de ese modo, apreciar la

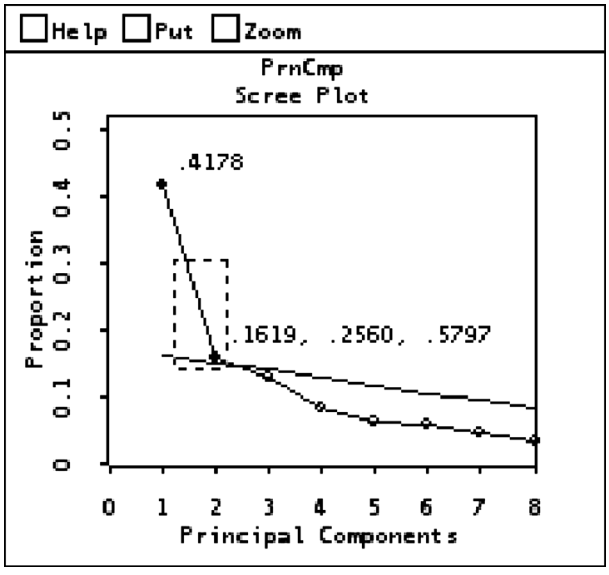

Figura 4. *Gráfico de sedimentación con línea de decisión basada en la propuesta de análisis paralelo de Kellie (2000)*

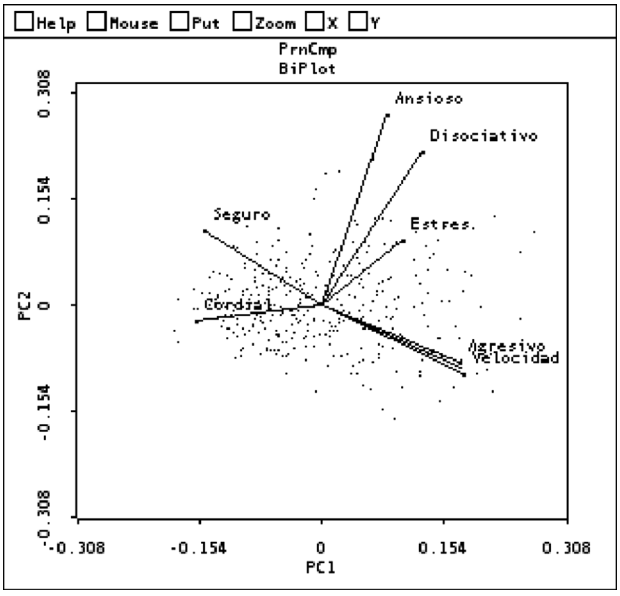

Figura 5. *Bi-plot sobre las dos primeros factores extraídos tras la aplicación del ACP*

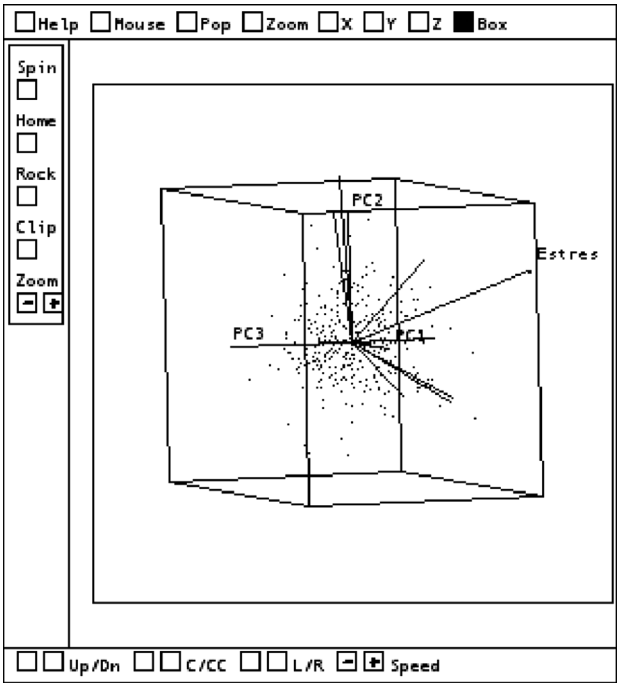

Figura 6. *Bi-plot tridimensional dinámico con las coordenadas de los casos —conductores— y variables —escalas del MDSI— en los tres primeros componentes principales*

pertinencia y significado de una solución trifactorial. Lógicamente, cuando la solución es bidimensional este gráfico es de poca utilidad, no obstante, puede seguir siendo interesante para explorar información residual a los dos primeros componentes. En nuestro caso, se observa que el tercer eje o componente está formado casi exclusivamente por el estilo *Reducción de Estrés* —que no aparece en la solución bidimensional—, aunque también en menor medida por el estilo *Cordial*. Esto sugiere la posibilidad de una relación entre ambos estilos que no quedaría representada en la solución en dos componentes. Mencionar también que, si bien en la representación gráfica implementada se presentan, por defecto, los tres primeros componentes extraídos, esto puede modificarse para representar cualquier subconjunto de componentes a través de los botones *X*, *Y*, *Z* de la barra de botones superior.

(4) Matriz de diagramas de dispersión con las coordenadas de las puntuaciones de los sujetos en los planos resultantes de la combinación por pares de los primeros componentes principales extraídos (figura 7) —el número de ellos dependerá de la resolución del monitor—. Aparte de ejercer de panel de selección de los componentes principales en el resto de los gráficos, esta representación gráfica va a permitir visualizar la posición de un sujeto o grupo de sujetos en varios factores simultáneamente. Así, si un grupo de 'puntos conductores' es seleccionado en el *Bi-plot* (figura 5), esta selección se mostrará inmediatamente en las nubes de puntos de todos los pares de componentes. También la detección de *outliers* puede verse facilitada a través de esta representación gráfica.

(5) Diagrama de puntos de las puntuaciones en los componentes principales extraídos (figura 8), el cual permite visualizar la distribución de las puntuaciones factoriales de los sujetos en cada uno de los componentes principales extraídos o, también, explorar perfiles individuales de puntuaciones a lo largo de los componentes —tal como se observa en la figura 8 para el caso del conductor '287' de la muestra—. En este gráfico se puede, opcionalmente, superponer un gráfico de caja y bigotes o un gráfico de diamantes (botones de la barra horizontal inferior).

Paso 3: Definir las técnicas de interacción de usuario

Al trabajar con gráficos interactivos es importante seleccionar los recursos que van a permitir al usuario interactuar con cada una de las representaciones gráficas y también con la VM en su conjunto. Estos elementos pueden aparecer de forma permanente en la VM, o bien mostrarse de forma puntual al ejecutarse determinadas acciones del usuario, por ejemplo, la aparición de un cuadro de diálogo demandando cierta información concreta. Como muestra de elemento permanente en nuestra VM se encuentra la lista con las etiquetas de las variables (escalas del MDSI en nuestro ejemplo) y observaciones (conductores) contenidas en el archivo de datos, la cual va a permitir identificar y seleccionar los casos y/o las variables en el resto de los gráficos de la VM (figura 9).

Destacar que el programa ViSta integra diferentes recursos de interacción para las representaciones gráficas ya disponibles por defecto, siendo algunos de estos comunes a diferentes tipos de gráficos. Así, por ejemplo, el etiquetado o «labeling» es una capacidad común a la mayoría de los gráficos de ViSta. No obstante, en ocasiones es necesario desarrollar nuevos recursos de interacción, como fue en el caso del gráfico de sedimentación de nuestra VM, el cual ha sido programado para admitir una forma especial de etiquetado: puesto que aquí lo que se representan no son casos sino los componentes principales, al seleccionar un punto de esta representación gráfica lo que se proporciona es información asociada a dicho componente, en concreto, la proporción de varianza ex-

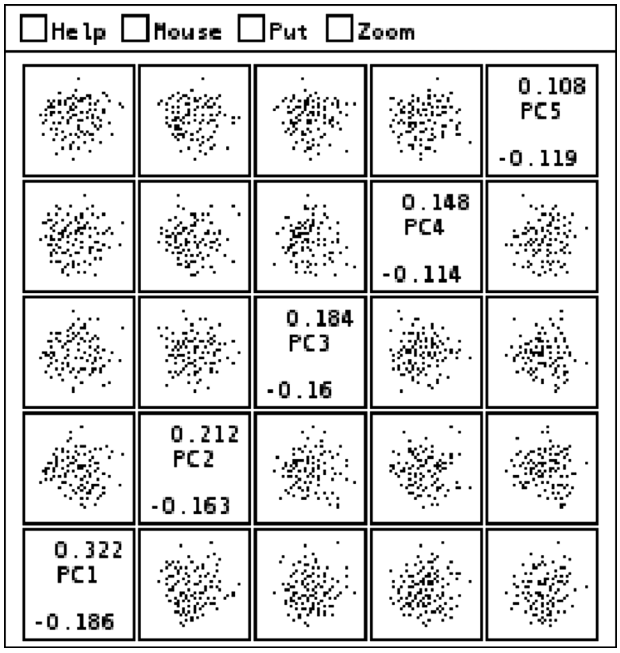

Figura 7. *Matriz de diagramas de dispersión con las puntuaciones de los sujetos —conductores de la muestra— en la combinación por pares de los cinco primeros componentes principales extraídos*

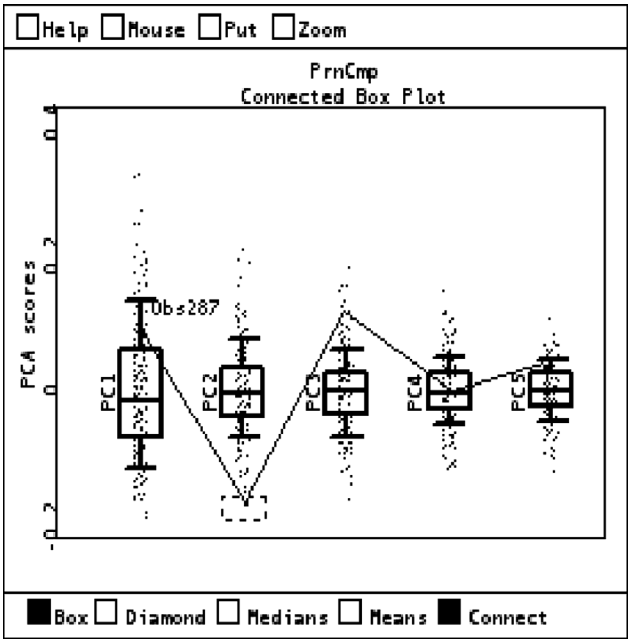

Figura 8. *Gráfico de diagrama de puntos y de caja y bigotes de las puntuaciones factoriales en los cinco primeros componentes extraídos*

plicada, la diferencia de proporción de varianza explicada en relación al componente previamente extraído, y la proporción de varianza explicada acumulada (véase figura 4 para el segundo componente extraído).

Paso 4: Definir las relaciones entre gráficos

Es necesario definir qué vínculos concretos van a ligar a unos gráficos con otros dentro de la VM, de modo que cambios en alguno de los elementos interactivos de la VM representen una actualización de la información mostrada en las representaciones gráficas en que ello resulte apropiado. Ello supone delimitar los canales de comunicación y los mensajes que van a circular entre los gráficos y los elementos interactivos de una VM, algo que no es del todo sencillo pues deben tenerse en cuenta las consecuencias de todas las posibles acciones del usuario cuando interactúe con la VM. Estos vínculos proporcionan a la visualización un sentido de conjunto, convirtiéndola en algo más que una colección de gráficos independientes.

En el caso de la VM implementada para el ACP existen vínculos de diferente tipo. Por ejemplo, una relación básica se establece a nivel de las observaciones: al seleccionar un caso o conjunto de casos —bien directamente en cualquiera de los gráficos de la VM, bien en el listado de todos ellos y de las variables que se integra en la parte izquierda de la VM—, esta selección se extiende al resto de los gráficos, es decir, los gráficos *comunican* la operación empírica realizada y resaltan las observaciones seleccionadas. En el ejemplo de la figura 9 podemos observar cómo la etiqueta «Obs287» ha sido marcada y, en consecuencia, ésta aparece resaltada en todos los gráficos en que el sujeto aparece representado.

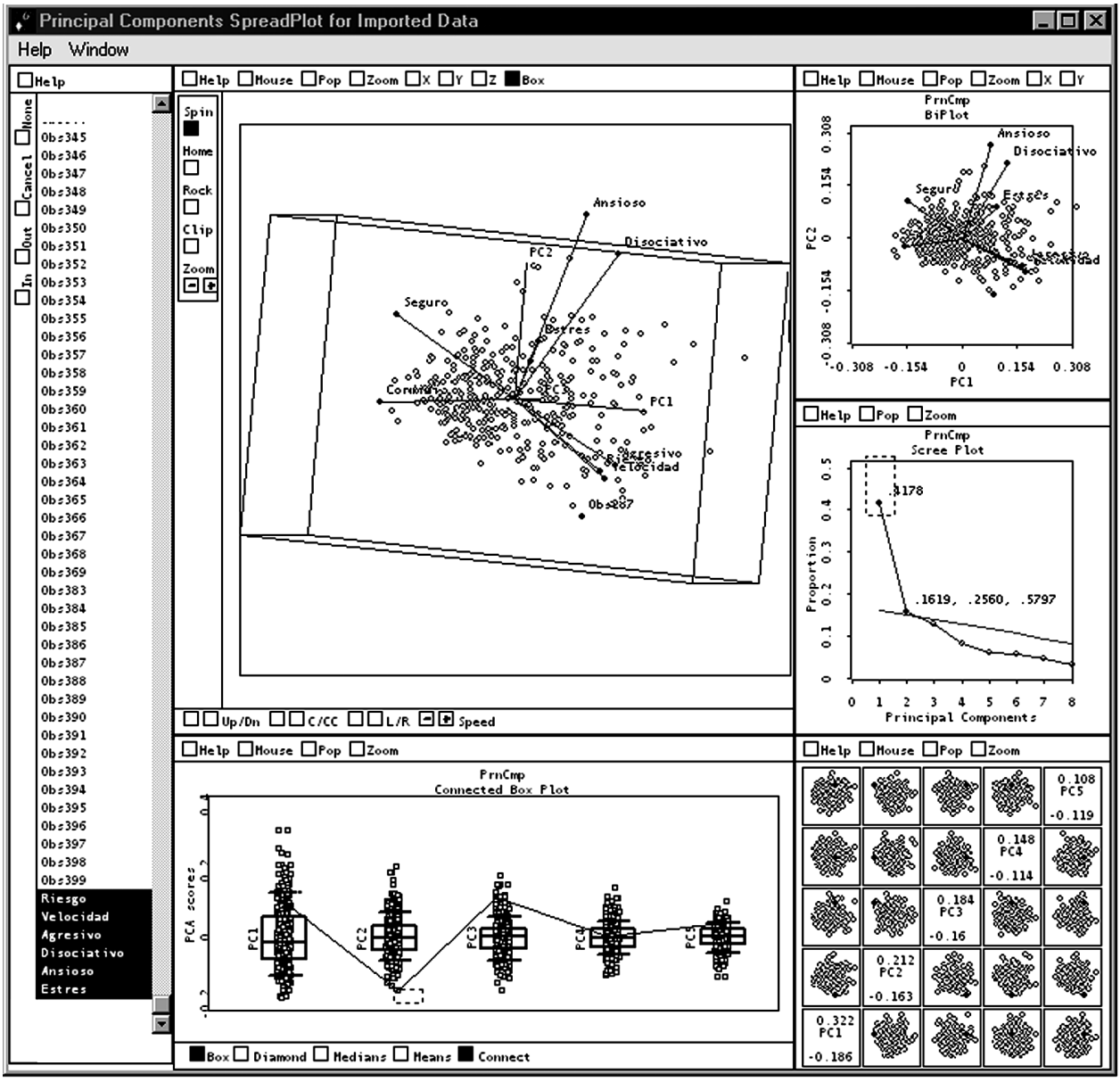

Figura 9. *VM en ViSta para el análisis de componentes principales*

Otro ejemplo de vínculo se da cuando se interactúa con la matriz de diagramas de dispersión: esta representación gráfica puede ser utilizada para seleccionar los componentes principales a representar en el resto de los gráficos de la VM —excepto en el gráfico de sedimentación—. Para ello, hay que hacer clic sobre las celdas de la diagonal de la matriz que correspondan a los componentes que deseemos seleccionar, lo cual permite explorar fácilmente los diferentes planos factoriales resultantes de combinar distintos componentes.

Paso 5. Definir la disposición de los gráficos en la visualización

Implica decidir la ubicación, tamaño e independencia relativa de cada una de las representaciones gráficas en la VM. La elección de estas opciones va a determinar la apariencia final de la VM y, en último término, la facilidad y eficiencia de la interacción de los usuarios con la misma. En la VM para el ACP (figura 9), aunque los gráficos son independientes y cada uno puede ser modificado en su tamaño y ubicación (botones *Zoom* y *Pop*, respectivamente), la visualización fue programada para que, por defecto, el diagrama 3D aparezca con un tamaño mayor en relación a los demás gráficos, motivado por la dificultad de interactuar con este tipo de gráfico si se presenta en un formato reducido. Por su parte, el gráfico con el diagrama de puntos aparece extendido horizontalmente a fin de facilitar la representación de un número amplio de variables (componentes, en este caso), del mismo modo que la ventana con el listado de etiquetas se ha extendido en su longitud y reducido en su anchura a fin de ajustarse mejor al tipo de contenido mostrado en la misma.

Paso 6: Definir la forma en que el programa interactúa con la VM

Implica programar la/s maneras/s en que el sistema en que se integre la VM interactuará con la misma, por ejemplo, cuándo podrá el usuario invocar la VM en el entorno de trabajo, cómo podrá hacerlo… En el caso de ViSta, existen varias formas en que el usuario puede activar la presentación de una VM, por supuesto, siempre y cuando se haya ejecutado el análisis que dé lugar a los resultados a partir de los que se pueda generar dicha VM: una forma es a través de la ventana de comandos («*Listener*» en ViSta), escribiendo un comando de visualización dirigido al objeto estadístico activo («send current-model: visualize»); otra, más fácil, desde la barra de menús del programa, seleccionando el comando *«Visualize Model»* del menú «Model».

Discusión

La VM como técnica gráfica ofrece una serie de posibilidades en el análisis de datos que descansan, básicamente, en su concepción como visualización global, dinámica e interactiva. Ello aporta la capacidad de crear formas de comunicación más intuitivas e interactivas entre el analista y los modelos estadísticos que aplique, si bien serán estudios empíricos y el «feedback» de los propios usuarios el que aporte la evidencia que permita evaluar la bondad de esta técnica gráfica.

Como se ha puesto de manifiesto a lo largo de este trabajo, ViSta proporciona una plataforma adecuada para la implementación de esta técnica gráfica, algo que actualmente no se puede afirmar para la mayoría de los sistemas estadísticos disponibles, muchos de los cuales ofrecen nulas o escasas posibilidades para el desarrollo de nuevos métodos gráficos, más aún si se trata de visualizaciones que impliquen capacidades dinámicas. Esta ventaja de ViSta tiene una parte importante de su origen en Lisp-Stat (Tierney, 1990), el lenguaje de programación con el que ha sido desarrollado este sistema, el cual se caracteriza por ser especialmente potente a nivel gráfico. Aún en los más potentes lenguajes de programación estadística disponibles en la actualidad, como es el caso de S-Plus (Insightful Co., 2006) y *R* (R Development Core Team, 2006), se carece de recursos eficientes a la hora de crear vínculos dinámicos entre gráficos, un aspecto fundamental en el diseño de una VM. Para una discusión detallada sobre Lisp-Stat, sus posibilidades, limitaciones y perspectivas en relación a otros lenguajes de programación estadística, sobre todo *R*, recomendamos la lectura del monográfico del *Journal of Statistical Software* sobre Lisp-Stat (Valero-Mora y Udina, 2005).

Ahora bien, el desarrollo de VMs no está exento de ciertas dificultades. Así, un primer problema asociado a la creación de una VM hace referencia a una limitación extrínseca al diseño de la misma: el tamaño limitado del medio físico en que va a ser visualizada, normalmente un monitor. Aunque es cierto que se cuenta con monitores cada vez más generosos en sus dimensiones y definición, la bondad de una VM implica un compromiso entre la cantidad de información que quiera incluirse en la misma y el número de gráficos que se pueden llegar a presentar simultáneamente en pantalla sin afectar la ergonomía visual de esta técnica gráfica.

Otro problema de las VMs hace referencia a la maleabilidad de su diseño. Así, las VMs disponibles en ViSta responden al diseño de quien en su momento eligió la información a ser mostrada, los gráficos que se incluirían en la misma, su disposición en pantalla, tamaño, etc. De hecho, en nuestro ejemplo de VM para el ACP, muy probablemente, otros analistas habrían optado por otro diseño de la VM en función de sus preferencias o modo habitual de operar en la ejecución de este tipo de análisis. El hecho es que el usuario del sistema ViSta no puede fácilmente modificar las visualizaciones ya existentes, salvo que se introduzca en los detalles de la creación de un «spreadplot» (VM) en ViSta (Young et al., 2003), algo que es posible dado el carácter abierto de este sistema pero que requiere un esfuerzo a tener en cuenta. A este respecto, el desarrollo de recursos que permitiesen al usuario redefinir de modo sencillo algunos aspectos de la configuración de una VM ya existente constituiría una vía importante de desarrollo en cuanto que otorgaría una gran flexibilidad a los analistas a la hora de contar con visualizaciones ajustadas a sus necesidades y preferencias.

Agradecimientos

Este trabajo ha sido realizado con el apoyo financiero de la Universidad Nacional de Mar del Plata y de la SeCyT (Argentina), así como de la Universitat de Valencia a través de la Fundación Sud-Nort y del programa para el profesorado para la realización de estancias cortas en otras universidades.

Referencias

Becker, R.A., y Cleveland, W.S. (1987). Brushing scatterplots. *Technometrics, 29*, 127-142.

Ben-Ari, O., Mikulincer, M., y Gillath, O. (2004) Themultidimensional driving style inventory-scale construct and validation. *Accident Analysis and Prevention*, *36*, 323-332.

Cattell, R.B. (1966). The scree test of the number of significant factors. *Multivariate Behavioral Research, 1,* 140-161.

Cleveland, W.S. (1993). *Visualizing data.* Murray Hill, NJ: AT&T Bell Lab.

Cleveland, W.S., y McGill, M.E. (1988). *Dynamic graphics for statistics*. Belmont, CA: Wadsworth.

DataDescription Inc. (2005). *DataDesk,* v. 5 [programa informático]. Disponible en: http.//www.datadesk.com.

Friedman, J.H., y Stuetzle, W. (2002) John W. Tukey's work on interactive graphics. *The Annals of Statistics*, *30*, 1629-1639.

Friendly, M. (2006). *Milestones in the history of thematic cartography, statistical graphics and data visualization.* Disponible en: URL: http://www.math.yorku.ca/SCS/Gallery/milestone/

Gabriel, K.R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. En V. Barnet (ed.): *Interpreting multivariate data* (pp. 147-174). Chichester: Willey.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185.

Insightful Co. (2006). *S-Plus*, v. 7 [programa informático]. Disponible en: http://www.insightful.com.

Keeling, K. (2000). A regression equation for determining the dimensionality of data. *Multivariate Behavioral Research, 35,* 457-468.

Molina, J.G., Ledesma, R., Valero-Mora, P., y Young, F.W. (2005). A Video Tour through ViSta 6.4, a Visual Statistical System based on Lisp-Stat. *Journal of Statistical Software, 13*(8), 1-10.

R Development Core Team (2006). *The R Project for Statistical Computing,* v. 2.2 [programa informático]. Disponible en: http://www.r-project.org/

Rivas, T., y Martínez Arias, R. (1991). Relación entre escalamiento multidimensional métrico y análisis de componentes principales *Psicothema, 3*, 443-451.

Ruiz, M.A., y San Martín, R. (1992). Una simulación sobre el comportamiento de la regla K1 para la estimación del número de factores. *Psicothema, 2,* 543-550.

Spss Inc. (2006) *SPSS 14.0 Base user's guide*. NJ: Prentice Hall.

StatSoft Inc. (2005). *Statistica Base* [programa informático]. Disponible en: http://www.statsoft.com.

Swayne, D., Lang, D.T., Buja, A., y Cook, D. (2003) GGOBI: Evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, *43,* 423-444.

Tierney, L. (1990). *Lisp-Stat An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. NY: John Wiley & Sons.

Tukey, J.K. (1977). *Exploratory data analysis*. MA: Addison-Wesley.

Tukey, J.K. (1980). We need both exploratory and confirmatory. *American Statistician*, *34,* 23-25.

Valero-Mora, P., y Udina, F. (eds.) (2005). Special volume: Lisp-Stat, Past, Present and Future. *Journal of Statistical Software, 13.* Disponible en: http://www.jstatsoft.org.

Wainer, H., y Velleman, P. (2001). Statistical graphics: Mapping the pathways of science. *Annual Review of Psychology, 52,* 305-35.

Young, F.W. (2006). *ViSta «The Visual Statistics System»* [programa informático]. Disponible en: http://www.visualstats.org.

Young, F.W., Faldowski, R.A., y McFarlane, M.M. (1993). Multivariate statistical visualization in computational statistics. En C.R. Rao (ed.): *Handbook of Statistics, 9,* 959-998. Amsterdam: Elsevier Science.

Young, F.W., Valero-Mora, P., y Friendly, M. (2006). *Visual statistics. Seeing data with dynamic interactive graphics.* NJ: Wiley & Sons.

Young, F.W., Valero-Mora, P., Faldowski, R., y Bann, C.M. (2003). Gossip: The arquitecture of spreadplots. *Journal of Computational and Graphical Statistics*, *12,* 80-100.