Metodología de Encuestas
Vol 3, Núm 1, 2001, 127-133

VISTA "THE VISUAL STATISTICS SYSTEM"

Forrest W. Young*
Rubén Ledesma Mouriño**
Gabriel Molina Ibáñez***
Noelia Llorens Aleixandre***
Pedro M. Valero Mora***
**University of North Carolina at Chapel Hill*
***Universidad de Mar del Plata*
****Universitat de València*

Introducción

ViSta "The Visual Statistics System" es un programa desarrollado por Forrest W. Young, profesor en el L. L. Thurstone Psychometric Laboratory de la Universidad de Carolina del Norte (EE.UU.). Este sistema informático forma parte del programa de investigación en métodos de visualización estadística, "The Visual Statistic Project" dirigido por el autor citado.

ViSta está inspirado en el enfoque del Análisis Exploratorio de Datos, quedando la *filosofía* del programa recogida en la frase ".mirando los datos para ver qué parecen decir" (Tukey, 1977). Este planteamiento se traduce en una diferencia importante con respecto a los programas estadísticos convencionales: ViSta enfatiza la utilización de recursos visuales incorporando métodos gráficos innovadores, métodos que se aplican tanto a la gestión, como a la transformación y al análisis de los datos.

Uno de los pilares del programa es la utilización de gráficos dinámicos e interactivos. A diferencia de los gráficos estadísticos convencionales, que son imágenes estáticas de los datos, los gráficos dinámicos pueden cambiar o transformarse, brindando diferentes imágenes de la estructura de los datos. Esos cambios pueden ser movimientos, animaciones o variaciones en tiempo real producidas por algún tipo de manipulación directa del usuario sobre el mismo gráfico u otros gráficos relacionados con éste. Las manipulaciones se ejecutan mediante acciones sencillas —por ejemplo, un movimiento del ratón— y producen una respuesta gráfica inmediata.

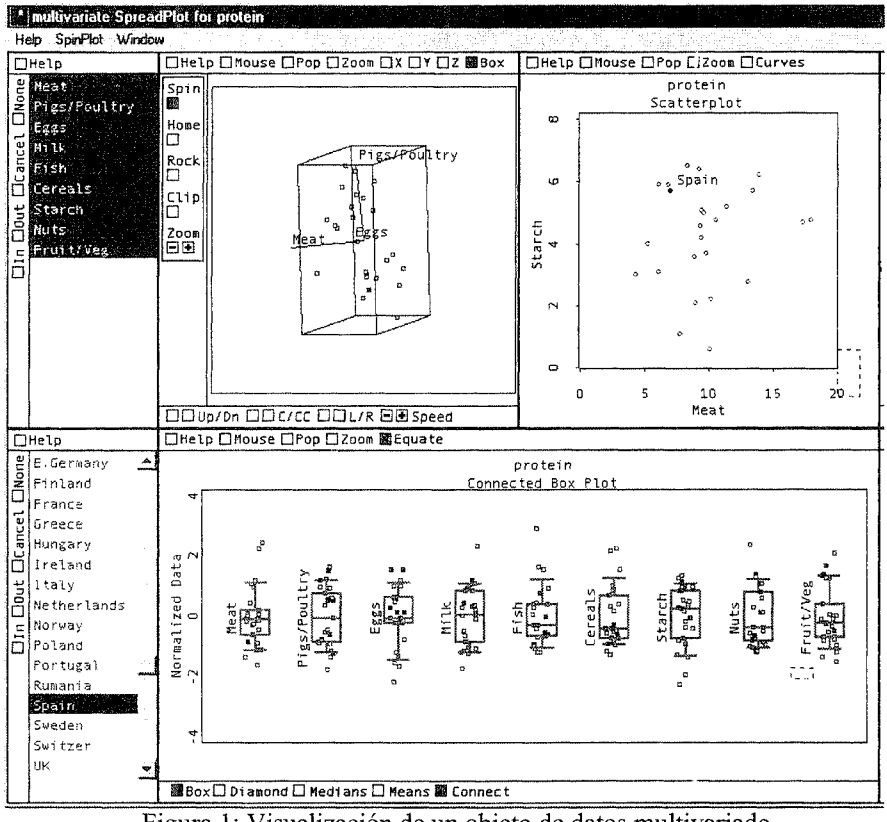

Figura 1: Visualización de un objeto de datos multivariado.

La permeabilidad a la acción del usuario confiere a esos métodos mayor capacidad para trasmitir información sobre los datos y mejor prestación en el análisis exploratorio de carácter multivariado. Entre las posibilidades de los gráficos dinámicos destaca el *ligado*, la posibilidad de vincular un gráfico con otro mediante relaciones empíricas, estadísticas, lógicas, algebraicas o definidas por el usuario. ViSta maximiza esa propiedad mediante los "spreadplots" : matrices de gráficos dinámicos vinculados entre sí, diseñadas para visualizar un tipo particular de objeto de datos, transformación o modelo estadístico. A modo de ejemplo, la Figura 1 muestra el *gráfico extendido* ("spreadplot") correspondiente a la visualización de un objeto de datos multivariado.

Interfaz gráfico

En la interacción con el usuario, ViSta posee una serie de recursos entre los que destacan los siguientes elementos:

- El *mapa de trabajo* (*"WorkMap")*: Se trata de un recurso gráfico que permite visualizar y controlar sesiones de análisis de datos. El mapa graba los pasos del analista y los representa como una secuencia de iconos conectados entre si por líneas. El usuario puede visualizar el proceso e interactuar con los pasos de la sesión, generando nuevos pasos y visitando o modificando pasos anteriores. La ventana del "WorkMap" también contiene una barra de herramientas con botones que permiten la aplicación directa de determinados modelos estadísticos.
- *Sistema de ayuda.* ViSta ofrece un amplio sistema de ayuda que cubre desde aspectos específicos a cuestiones más globales, relativas al funcionamiento general del programa. Diferentes funciones permiten acceder a toda la información de ayuda disponible.
- *Otros*: un editor de datos en formato de hoja de cálculo (*"DataSheet"*) que permite introducir y visualizar los datos de nuestros archivos; un panel de control de variables y observaciones (*Selector*) en el que poder seleccionar aquéllas que resulten de interés en un momento dado; y la ventana *Listener* como herramienta más orientada a usuarios avanzados en cuanto que permite introducir y ejecutar comandos en lenguaje Lisp-Stat, a la vez que controlar la ejecución del programa.

Transformación de datos

En ViSta pueden diferenciarse tres categorías de posibles transformaciones que pueden ser aplicadas sobre los datos objeto de análisis:

- *Transformaciones predefinidas.* Son transformaciones usualmente incluidas en los paquetes estadísticos, como la creación de rangos, la normalización de variables, las transformaciones logarítmicas, trigonométricas, etc.
- *Transformaciones definidas por el usuario.* El editor de Lisp-Stat del ViSta permite programar al usuario sus propias transformaciones. Adicionalmente, ViSta incluye un lenguaje propio, el ViVa ("ViSta's Interactive Variable"), que permite construir expresiones algebraicas de manera más específica y sencilla que con en Lisp-Stat.
- *Transformaciones visuales.* Son transformaciones estadísticas gráficas de carácter interactivo y dinámico. Permiten al usuario seleccionar los mejores parámetros para una determinada transformación, utilizando para ello el recurso visual de los gráficos extendidos. Quedan incluidas en esta categoría

las transformaciones de Box Cox y las denominadas de potencias plegadas ("folder-power transformations").

Métodos estadísticos

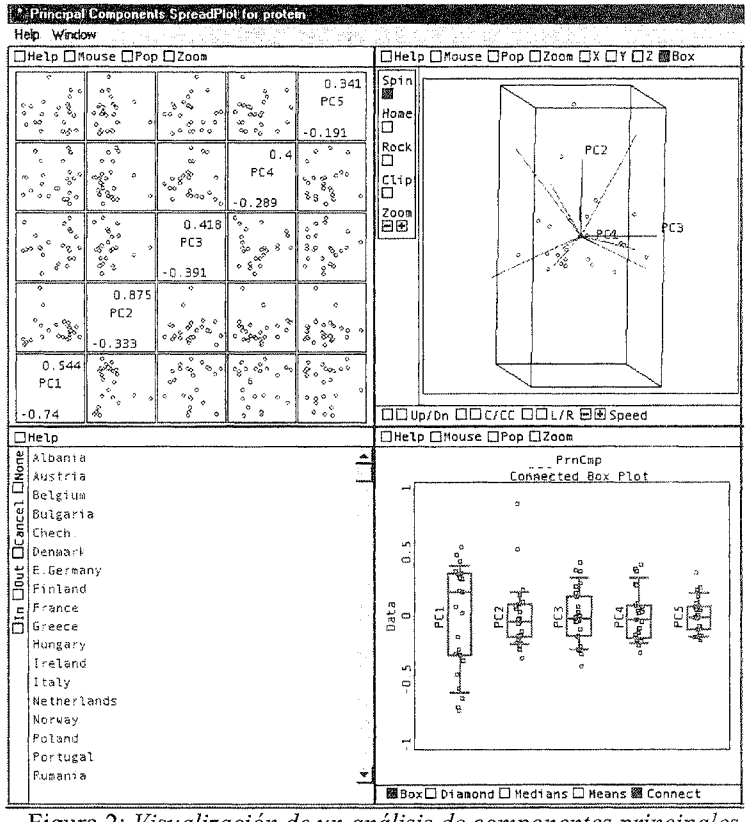

Figura 2: *Visualización de un análisis de componentes principales.*

ViSta incluye una gama amplia de opciones para el análisis estadístico de datos que, además, no deja de crecer dada la arquitectura abierta de este programa. En la actualidad se encuentran disponibles los siguientes procedimientos de análisis:

- Tests estadísticos univariados (T, Z, etc.).
- Análisis de varianza.
- Análisis de regresión (OLS, Robusta, Monótona).
- Regresión multivariada.
- Análisis de tablas de frecuencias.
- Análisis de correspondencias y de homogeneidad (correspondencias múltiples).

- Escalamiento multidimensional.
- Análisis cluster.
- Análisis de componentes principales.
- Exploración y tratamiento de datos ausentes.
- Análisis log-lineal.

En ViSta, los resultados de los análisis aplicados pueden ser presentados en formato tradicional ("output" de texto) o visualizados mediante gráficos extendidos. Un ejemplo de cada uno de estos dos formatos de presentación de los resultados en ViSta se muestra en las Figuras 2 y 3, ambos obtenidos tras la aplicación de un análisis de componentes principales sobre un mismo archivo de datos. Asimismo, los resultados del análisis pueden ser utilizados para generar nuevos archivos de datos y realizar análisis adicionales.

Características técnicas

ViSta es un programa gratuito que puede obtenerse vía Internet en la dirección www.visualxtats.org y, probablemente de forma más rápida, a través del "mirror" de esta página en España (www.uv.es/~prodat/ViSta). Funciona bajo diferentes plataformas (Windows, Macintosh y Unix) y está disponible en inglés (en todas su versiones), en francés y en castellano. Se trata de un programa escrito en el lenguaje de programación Lisp-Stat (Tierney, 1990), lenguaje concebido principalmente como una herramienta para la investigación y el desarrollo en métodos de visualización estadística. ViSta ofrece la posibilidad de adaptar los procedimientos estadísticos y de visualización disponibles a las necesidades de cada usuario y, por supuesto, crear e incorporar nuevos de ellos al programa. Puede obtenerse más información sobre esta última posibilidad en la sección de la página web del ViSta orientada a desarrolladores.

Comentarios finales

En cuanto a la situación actual y el desarrollo futuro del programa, su proyección dentro de la oferta de paquetes estadísticos radica en el desarrollo de métodos novedosos y recursos no disponibles en programas estadísticos convencionales. En ese sentido, el entorno está abierto a múltiples posibilidades de innovación, lo que le supone además un atractivo para el investigador interesado en la materia.

Es de esperar que en el futuro se incremente la utilización de métodos visuales en el análisis de datos, reconociendo la utilidad comparativa de esos recursos en ciertas tareas y situaciones de análisis. En tal sentido, y puesto que se trata de métodos emergentes, aún restan por explorar muchas de sus potencialidades. Ello supone una buena perspectiva en cuanto a la posibilidad de innovación y desarrollo de nuevos procedimientos para datos, modelos o tareas específicas.

```
Principal Components Analysis of Variable Correlation

Model:     PrnCmp
Variables: (Meat Pigs/Poultry Eggs Milk Fish Cereals Starch Nuts Fruit/Veg)

Correlation Matrix
                    VARIABLES
VARIABLES           Meat Pigs/Poult      Eggs      Milk      Fish    Cereals
Meat              1.0000    0.1530    0.5856    0.5029    0.0610   -0.4999
Pigs/Poultry      0.1530    1.0000    0.6204    0.2815   -0.2340   -0.4138
Eggs              0.5856    0.6204    1.0000    0.5755    0.0656   -0.7124
Milk              0.5029    0.2815    0.5755    1.0000    0.1379   -0.5927
Fish              0.0610   -0.2340    0.0656    0.1379    1.0000   -0.5242
Cereals          -0.4999   -0.4138   -0.7124   -0.5927   -0.5242    1.0000
Starch            0.1354    0.3138    0.4522    0.2224    0.4039   -0.5333
Nuts             -0.3494   -0.6350   -0.5598   -0.6211   -0.1472    0.6510
Fruit/Veg        -0.0742   -0.0613   -0.0455   -0.4084    0.2661    0.0469

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               4.00644    0.44516     0.44516
PC2               1.63500    0.18167     0.62683
PC3               1.12792    0.12532     0.75215
PC4               0.95466    0.10607     0.85822
PC5               0.46384    0.05154     0.90976
PC6               0.32513    0.03613     0.94589
PC7               0.27161    0.03018     0.97607
PC8               0.11629    0.01292     0.98899
PC9               0.09911    0.01101     1.00000

Coefficients (Eigenvectors):
                    COMPONENTS
VARIABLES             PC1         PC2         PC3         PC4         PC5         PC6
Meat               0.3026     -0.0563      0.2976      0.6465      0.3222      0.4599
Pigs/Poultry       0.3106     -0.2369     -0.6239     -0.0370     -0.3002      0.1210
Eggs               0.4267     -0.0353     -0.1815      0.3132      0.0791     -0.3612
```

Figura 3: *Parte de la salida de texto en un análisis de componentes principales.*

También es de suponer que el interés comercial de los paquetes estadísticos convencionales, especialmente en lo que respecta al mejoramiento de la interfaz con el usuario, impulse una paulatina incorporación de gráficos dinámicos como parte de los recursos ofrecidos. No obstante, es poco probable que la filosofía de la Visualización y el Análisis Exploratorio de Datos como se presenta en algunos programas comerciales (DataDesk) y no comerciales (ViSta), pueda ser incorporada en toda su amplitud por los paquetes estadísticos convencionales, pues ello supondría un cambio radical en la estructura de los mismos.

Referencias

Becker, R. A., Cleveland, W. S. y Wilks, A. R. (1987). Dynamic Graphics for Data Analysis. *Statistical Science, 2*, 55-295.

Chambers, J. M., Cleveland, W. S., Kleiner, B. y Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.

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.

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

Tukey, J. K (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

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

Young, F. W., Faldowski. R. A. y McFarlane, M. M. (1993). Multivariate Statistical Visualization. En Rao, C. R. (Ed.): *Handbook of Statistics, Vol 9.* (959-998). Amsterdam: Elsevier Science.

Young, F. W. y Rheingans, F. (1991). Visualization Structure in High-Dimensional multivariate Data. *IBM Journal of Research and Development, 35*, 97-107.

Young, F. W. y Lubinsky, D. J. (1995) Guiding Data Analysts with Visual Statistical Strategies. *Journal of Computational and Graphical Statistics*. 4, 229-250.

Young, F. W., Valero, P. M. (2000). *Transformations*. Report Number 2000-3. L. L. Thurstone Psychometric Lab.: Univ. of North Carolina at Chapel Hill, USA.

Young, F. W., Valero, P. M., Faldowsky, R. A. y Bann, C. (2000). *SpreadPlots*. Report Number 2000-4. L. L. Thurstone Psychometric Lab.: Univ. of North Carolina at Chapel Hill, USA.