Metodología de Encuestas
Vol 6, Núm 1, 2004, 71-75
ISSN: 1575-7803

INTERFACES GRÁFICOS DE USUARIO PARA ANÁLISIS DE DATOS

Pedro M. Valero
Universitat de Valéncia

RESUMEN

Hace 40 años Tukey (1965) señaló que era el momento para que los estadísticos utilizaran los ordenadores de una manera "flexible, fluida y productiva". En aquel momento, se estaban desarrollando buena parte de los paquetes o entornos de programación estadística que todavía hoy en día seguimos utilizando. Estos programas estaban basados en una interfaz de lenguaje de comandos que, aunque el más adecuado para la tecnología disponible en aquel momento, hoy puede no resultar tan satisfactorio para todas las tareas y usuarios.

Palabras clave: *análisis de datos, interfaces gráficos, ViSta.*

Introducción

Hace ahora aproximadamente 40 años, Tukey (1965) señaló que había llegado el momento en que los estadísticos deberían empezar a utilizar los ordenadores para el análisis de datos una manera flexible, fluida y productiva. Ahora bien, para lograr ese objetivo, los programas y entornos de cálculo deberían tener las características adecuadas para que fuera posible interactuar con ellos de esa manera. Puesto que en aquel momento, y durante al menos la década siguiente, la única forma realista de interactuar con el ordenador fue utilizar un lenguaje de comandos. Flexibilidad, fluidez o cantidad de resultados dependerían básicamente de lo bien desarrollado y estructurado que estuviera ese lenguaje.

Ahora bien, los lenguajes de comandos no son la mejor opción para todos los usuarios y todas las tareas. Así, en cuanto a los usuarios, Chambers (1999) señala que los lenguajes de comandos no son la solución más apropiada para "el ejecutivo ocupado, el no especialista de otro campo, la persona no técnica que precisamente necesitan respuestas muy sencillas". Por otro lado, parece natural que tareas tales como la visualización dinámica de datos, en las que el usuario es capaz de interactuar con la representación para extraer elementos de interés, serán realizadas mejor por medio de un estilo de interacción visual, también denominado de manipulación directa, en lugar de por medio de un estilo de interacción verbal, tal y como el requerido por los lenguajes de comandos.

A mediados de los años 70, una serie de pioneros, entre los que se encontraba, naturalmente, Tukey, dieron los primeros pasos en relación con la aplicación de métodos de manipulación directa a gráficos estadísticos (Fisherkeller y col, 1975). Estos experimentos fueron uno de los antecedentes más importantes de los denominados gráficos dinámicos (Cleveland y McGill, 1988). Pronto, una serie de técnicas tal y como el ligado (Newton, 1978; Stuetzle, 1987), el cepillado (Becker y Cleveland, 1987) o la selección e identificación (Wills, 1996) fueron implementados con objeto de ser evaluados en cuanto a su utilidad para el análisis de datos. Al principio, esos programas fueron en la mayor parte de los casos sistemas dirigidos a demostrar el concepto y no cubrían un número grande de situaciones de análisis. No obstante, a mediados de la década de los 80, ya aparecieron una serie de sistemas estadísticos que ofrecían una mayor variedad de técnicas y que de hecho podrían ser alternativas realistas a los programas comerciales de mayor popularidad en el momento. Entre estos programas tendríamos en el campo comercial DataDesk (Velleman y Velleman, 1985) y JMP, mientras que en el campo no comercial se encontrarían Lisp-Stat (Tierney, 1990), ViSta (Young, 1992), XGobi (Swayne, Cook y Buja, 1998) y MANET (Unwin y col, 1996).

Ahora bien, esos programas no sólo eran renovadores por incluir gráficos dinámicos sino porque los diseñadores de estos programas tuvieron que hacer un uso muy importante de los interfaces gráficos y de manipulación directa. Así el estilo flexible, fluido y en cantidad propugnado por Tukey se conseguía *por medio de un interfaz gráfico*, y para ello no bastaba con un lavado de cara tal y como ciertos programas comerciales realizaron en su momento, incorporando menús, cuadros de diálogo, etc. que, en realidad, simplemente llevaban a cabo la tarea de construir los comandos a ejecutar. Estos programas, por el contrario, llevaron a cabo un análisis profundo de las tareas estadísticas y de la

manera en que un sistema informático de estas características podrían ser implementadas. Conceptos tales como metáforas, iconos, acciones, etc. fueron utilizados de una manera planeada produciendo programas que, en algunos casos, ofrecían una experiencia de usuario de gran calidad.

La situación actual

La situación actual en relación con las herramientas para análisis de datos puede decirse que está sobre todo marcada por el crecimiento del uso del lenguaje de programación S y de su versión gratuita R (Becker y Chambers, 1985; Becker y col, 1988; Ihaka y Gentleman, 1996). Este sistema está teniendo un desarrollo que parece vertiginoso comparado con el que parece experimentar el software comercial. En la actualidad, es posible encontrar más métodos estadísticos, más avanzados y más completos simplemente por medio de su descarga gratuita que pagando un precio a menudo muy elevado por un programa comercial. El éxito de R ha sido tan importante que es posible observar un sentimiento de euforia creciente según el cual éste puede convertirse en una alternativa real al software comercial para todo tipo de tareas y todo tipo de usuarios.

No obstante, esta euforia debería moderarse si tenemos en cuenta que R está basado en un lenguaje de comandos el cual, aunque bien diseñado y que permite satisfacer los tres criterios de Tukey, es sobre todo adecuado, obviamente, para usuarios expertos que conocen el lenguaje. Los usuarios ocasionales, en cambio, resulta difícil que acepten alegremente los desafíos que el software no comercial suele imponer, por mínimos que sean, y, por tanto, preferirán evitarlos a pesar del coste económico que ello supone. ¿Y qué es lo que estos usuarios necesitan? Dicho en términos simples, lo que necesitan es un interfaz de usuario gráfico que les permita un grado similar de flexibilidad, fluidez y productividad similar a la que tienen los expertos que utilizan R pero sin tener que aprender un lenguaje de programación.

En la actualidad, existen varios intentos de proporcionar un interfaz gráfico a R aunque ninguno de ellos puede calificarse como muy satisfactorio (algunos de los que se mencionan en *http://www.sciviews.org/_rgui/* por ejemplo son R Commander, Brodgar, SciViews y ObveRsive). En este momento, en opinión del autor de este texto, si alguien está interesado en un programa gratuito para análisis de datos que tenga un interfaz gráfico relativamente completo, la mejor opción disponible es posiblemente ViSta, más aún si ese alguien está interesado en desarrollar nuevas técnicas de exploración e interacción gráfica con datos[1].

En cuanto al futuro deseable, es obvio que sería de gran interés que R incorporara herramientas que permitieran generar interfaces gráficos de calidad. No obstante, aquí nos encontramos con una serie de limitaciones técnicas bastante importantes. Mientras que, por ejemplo, Lisp-Stat era un lenguaje que incluía primitivas orientadas a la creación de interfaces de usuario, R no tiene esa capacidad. Para solucionarlo se han utilizado conexiones a TclTk, Python y, como no, Java, pero, en nuestra opinión, obligar a los desarrolladores a tener que aprender esos otros lenguajes no va a favorecer la proliferación de nuevos interfaces. Sin embargo, dado el grado de entusiasmo demostrado por la

[1] Pido perdón por la auto-propaganda.

comunidad científica hacia R, parece difícil que estas dificultades técnicas no se vean finalmente superadas.

Téngase en cuenta que el párrafo anterior indica *herramientas para la construcción de interfaces*, no un interfaz propiamente dicho. Idealmente, R debería servir en esta área para lo mismo que sirve en otras áreas. Favorecer la exploración de nuevas técnicas e ideas para el análisis de datos al permitir la implementación de éstas de una manera relativamente rápida y eficiente. Hay que advertir, sin embargo, que a menudo resulta mucho más simple en términos de esfuerzo y tiempo de programación ofrecer una nueva técnica o rutina estadística que un interfaz de calidad, y que el reconocimiento que un especialista en estadística puede obtener por lo primero es mucho mayor que por lo segundo. No existen por ejemplo muchas revistas científicas que aceptarían artículos sobre una nueva manera de manipular tablas de contingencia que permitiera, por ejemplo reagrupar categorías utilizando el ratón. La mejora de los recursos técnicos quizás debería acompañarse de un cambio en la forma en que son valorados este tipo de esfuerzos entre la comunidad científica.

Finalmente, aunque los responsables del software comercial están en una mejor situación para llevar a cabo este tipo de implementaciones, lo cierto es que a menudo carecen del tiempo o los recursos para llevar a cabo desarrollos conceptuales o teóricos que guíen las implementaciones concretas. Ello puede llevar a programar, a menudo con mucho esfuerzo, ideas que resultan sólo apropiadas para circunstancias muy concretas y que carecen de generalidad. En este sentido, apoyarse en expertos que hayan tenido la oportunidad de explorar las posibilidades de estos interfaces gráficos en relación con el análisis de datos puede resultarles de mucha utilidad.

El futuro

En mi opinión, el futuro es moderadamente optimista. R es un factor de renovación muy importante y existe una gran cantidad de entusiastas interesados en hacer que progrese. Si estos entusiastas consiguen desarrollar este programa de tal manera que sea posible construir interfaces de calidad nos encontraremos de repente ante un programa con una potencialidad impresionante. Los desarrolladores de software comercial, por su parte, tendrán que trabajar duro para superar este listón. En este contexto de competencia, todos deberíamos salir beneficiados.

Referencias

Becker, R. A. and Chambers, J. M. (1985). Design of the S system for data analysis. *Atandt Tech. J.*, 64 (9) 2131–2151.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The New S Programming Language*. Wadsworth and Brooksslash Cole, Pacific Grove, CA, USA and Pacific Grove, CA, USA.

Becker, R. A. and Cleveland, W. S. (1987). Brushing scatterplots. *Technometrics*, 29:127–142.

Chambers, J. (1999). Computing with data: Concepts and challenges. *American Statistician*, 53 (1) 73–84.

Cleveland, W. C. and McGill, M. E. (1988). *Dynamic Graphics for Statistics*. CRC Press, Inc.

Deborah F. Swayne, Dianne Cook and Buja, A. (1998). XGobi: Interactive dynamic data visualization in the X Window System. *Journal of Computational and Graphical Statistics*, 7 (1) 113–130.

Fisherkeller, M. A., Friedman, J. H., and Tukey, J. W. (1975). Prim9: a data display and analysis system. In *Pacific Regional Conference of the Association for Computing Machinery*, San Francisco, CA.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5 (3) 299–314.

Newton, C. M. (1978). Graphics: From alfa to omega in data analysis. In Wang, P. C. C., editor, *Proc. of the Symposium on Graphical Representation of Multivariate Data*, New York, NY. Academic Press.

Stuetzle, W. (1987). Plot windows. In Cleveland, W. S. and McGill, M. E., editors, *Dynamic Graphics for Statistics*, pages 225–246, Belmont, Ca. Wadsworth.

Tierney, L. (1990). LISP-STAT: an object oriented environment for statistical computing and dynamic graphics. Wiley-Interscience.

Tukey, J. W. (1965). The technical tools of statistics. *The American Statistician*, 19, 23–28.

Unwin, A., Hawkins, G., Hofmann, H., and Siegl, B. (1996). Interactive graphics for data sets with missing values - Manet. *Journal of Computational and Graphical Statistics*, 5 (2) 113–122.

Velleman, P. F. and Velleman, A. Y. (1985). *Data Desk*. Data Description Inc.

Wills, G. J. (1996). Selection: 524,288 ways to say 'this is interesting'. In *Proceedings of IEEE Symposium on Information Visualization (InfoVis '96)*, pages 54–60.

Young, F. (1992). ViSta: The Visual Statistics System. Technical report, UNC Psychometric Laboratory.