Coordinación de visualizaciones en diferentes ventanas: El caso de la representación gráfica de datos estadísticos

, Forrest W. Young[2], Pedro M. Valero Mora[1] y Ignacio Pareja Montoro

[1]Instituto de Tráfico y Seguridad Vial, Universitat de València, C/ Hugo de Moncada n, 2 Entresuelo. Valencia 46010. valerop@uv.es
[2]University of North Carolina at Chapel Hill. CB 3270 DA, Chapel Hill NC., USA 27599-3270. forrest@unc.edu

Resumen: El siguiente artículo revisa algunas de las lecciones que hemos aprendido en relación con el desarrollo de sistemas de visualización de datos estadísticos que utilizan varias vistas. El texto se acompaña de ejemplos tomados del programa estadísticoViSta [0] desarrollado principalmente por el primer autor y con colaboraciones del resto de los autores.

Introducción

Hay ocasiones en que la información por sí misma es suficiente para nuestros objetivos. Sin embargo, en la mayoría de las ocasiones, es necesaria una elaboración de esa información que nos permita obtener respuestas a las preguntas que nos hacemos. Las técnicas de visualización de la información pretenden llegar a representaciones que hagan más fácil esa elaboración. A partir de ellas, los usuarios son capaces de extraer conclusiones que de otro modo estarían escondidas tras la información en estado bruto. No obstante, estas representaciones no deberían ser estáticas sino que deberían posibilitar la posterior modificación de las representaciones, sugeridas por aquellas, que permitirían probar nuevas hipótesis, refinar las ya encontradas o, bien, poner a prueba lo ya averiguado. Existen en la actualidad una gran cantidad de esfuerzos en esa dirección como se puede comprobar en [1]. Por tanto, las representaciones visuales deben ser capaces de proporcionar soporte al proceso de exploración de la información y necesitan adaptarse a los nuevos requerimientos que puedan surgir *durante el proceso mismo.*

Una de las conclusiones que a menudo puede extraerse de las técnicas de visualización es que no existe una única representación que sea capaz de responder a todas las preguntas que se puedan plantear. Por el contrario, resulta muy conveniente observar varias visualizaciones de diferente tipo simultáneamente para combinarlas a la hora de comprender el problema en cuestión. Por ejemplo, en [2] se describe un ejemplo de un sistema multimedia que permite identificar complicaciones en niños durante el embarazo. Estas complicaciones se producen de una manera muy sutil y los médicos requieren a menudo revisar simultaneamente datos fisiológicos y la observación visual de los movimientos de los niños. Estos autores afirman que los diseñadores de sistemas de visiones múltiples a menudo cometen errores de diseño, introduciendo complejidades innecesarias e inconsistencias cuando intentan coordinar las diferentes visiones en el interface. Así, estos autores proporcionan recomendaciones para ayudar a los diseñadores a evitar estos errores.

Un problema también de interés es el afrontado por [3]. Ellos afirman que a menudo existen programas capaces de llevar a cabo un tipo de visualización pero no existe manera de ligar esa visualización con la producida por otro programa. Ellos desarrollan un sistema para coordinar visiones múltiples. Este sistema permite tomar ventaja de relaciones simples entre los datos de tal manera que la coordinación entre diferentes visiones pueda ser posible sin programación.

Un área dentro de la existe una cierta tradición de utilización de visiones múltiples es la de del análisis de datos estadísticos [4]. En esta área existen una gran cantidad de gráficos que pueden ser apropiados según el tipo de datos, la complejidad, las relaciones asumidas entre ellos, etc. Desde los años noventa ya se empieza a formalizar técnicas que permiten conectar los diferentes gráficos estadísticos de tal manera que podemos hablar de visiones múltiples interconectadas [4]. Desde entonces ha habido una gran cantidad de exploración en interfaces que mejoran esos primeros intentos. El siguiente artículo expone nuestro trabajo en esa dirección así como los principios de diseño que hemos ido desarrollando en relación con él.

El caso de las representaciones gráficas de datos estadísticos.

Introduciremos de una manera breve algunas de las representaciones gráficas de datos estadísticos. Esta descripción no pretende ser exhaustiva sino que sólo intenta ser un comienzo para la elaboración del resto del trabajo.
Para nuestros propósitos podemos distinguir entre representaciones gráficas simples y compuestas.

Representaciones simples.

Como un ejemplo, en la figura 1 es posible ver varios gráficos que pueden ser apropiados en función de que nos centremos en una dimensión, en dos dimensiones, en tres, o en más de tres. Los datos corresponden al precio de una hamburguesa en diferentes capitales del mundo junto con el precio de otros y a los datos de supervivencia en el Titanic en función del Sexo, la edad y la clase en la que viajaban.

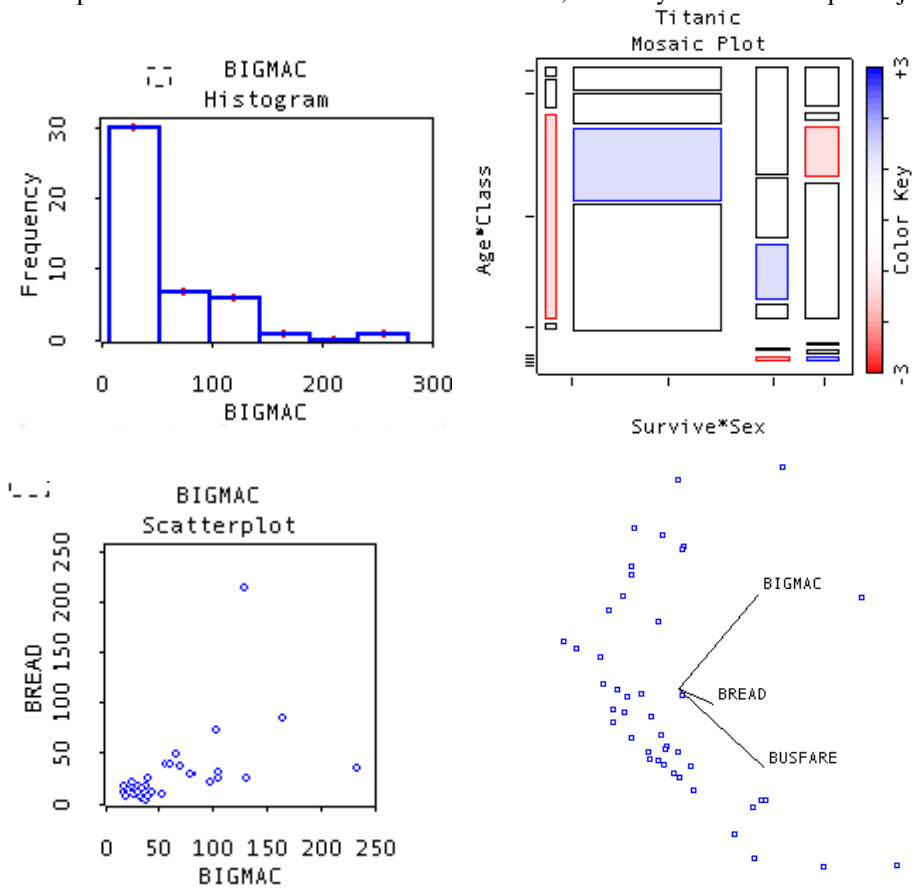

Figura 1: Gráficos estadísticos simples

Representaciones compuestas.

En las representaciones compuestas, varios gráficos estadísticos son repetidos para formar un nuevo gráfico. Un ejemplo es la matriz de diagramas de dispersión. En ella, el mismo gráfico es repetido varias veces para varias variables. De esta manera es posible tener una impresión de las relaciones entre las variables de un conjunto de datos. La diagonal es aprovechada para mostrar un gráfico univariado y los nombres de las variables en la fila/columna.

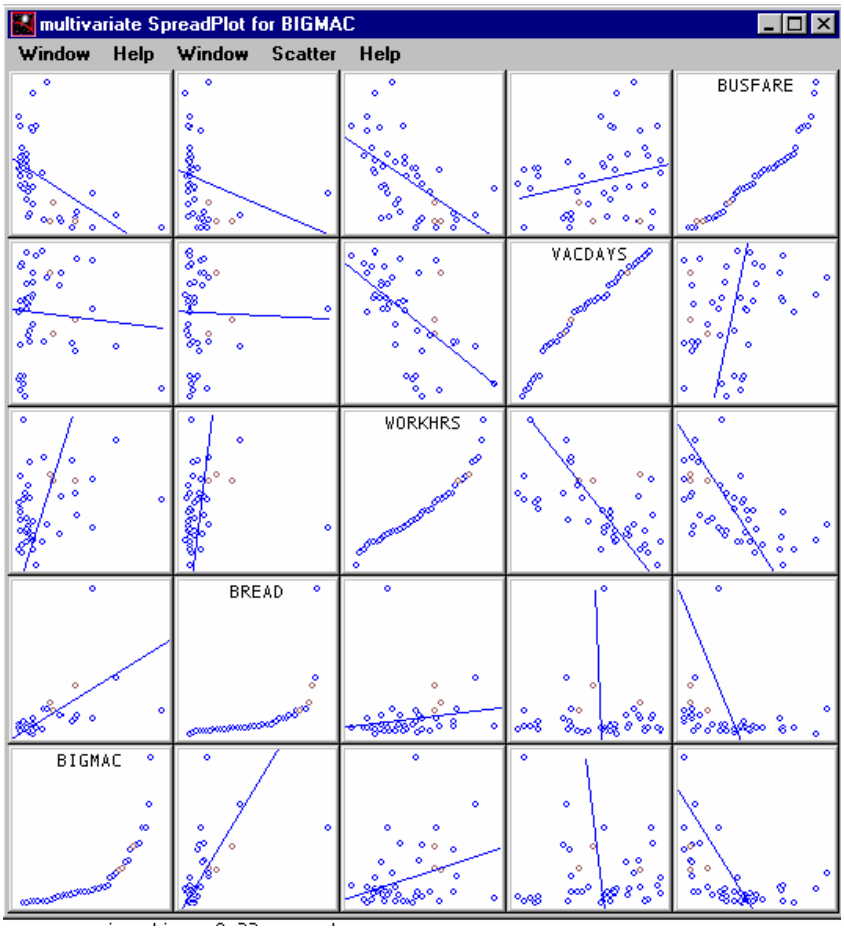

Figura 2: Matriz de diagramas de dispersión

Múltiples Visiones

Ya se trate de gráficos simples o compuestos lo cierto es que a menudo existe la necesidad de combinarlos y utilizar varios de ellos simultáneamente. Un ejemplo de estos puede observarse en la figura 3. Los datos a los que se aplica son los correspondientes a la supervivencia en el Titanic en función del género, la edad y la clase social de los pasajeros En este gráfico se representa un modelo estadístico que para su correcta interpretación necesita de varios gráficos a la vez. Por ejemplo, el tamaño de los rectángulos en el gráfico de mosaico (margen superior derecha) son proporcionales al recuento de los casos para el cruce de variables. Esa información debe ser contrastada con el ajuste global que se observa en la ventana flotante de texto y en el gráfico de puntos y líneas que permite la comparación con el ajuste global de otros modelos. La información del modelo se muestra en la ventana de la izquierda en la que los elementos seleccionados son los considerados de relevancia en él. En este caso se estaría probando un modelo según el cual la supervivencia no derivaría de la edad y del sexo de los viajeros ni de su edad y clase. Durante los dos últimos años hemos estado implicados en el desarrollo de gráficos de este tipo y ello nos ha llevado a una serie de consideraciones que discutiremos a continuación.

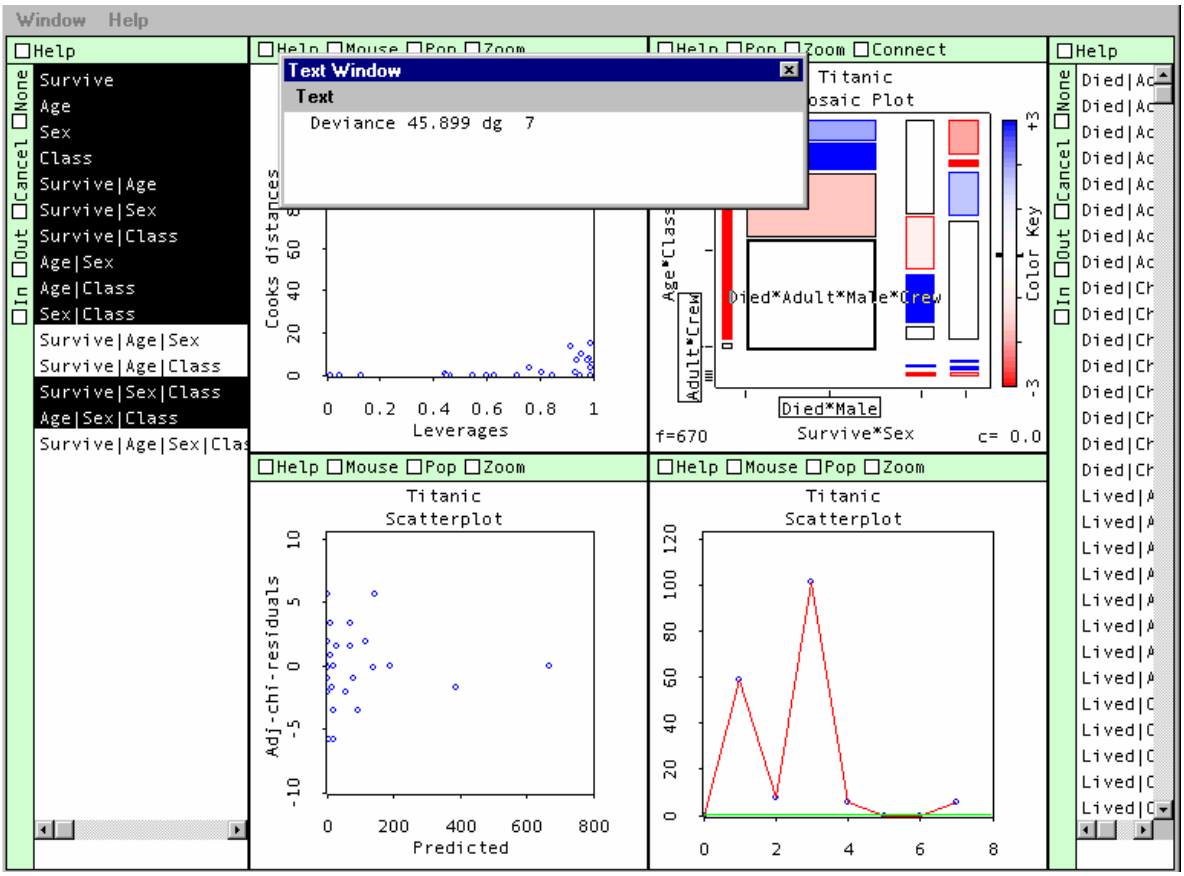

Figura 3: Gráfico estadístico compuesto de gráficos simples

Consideraciones sobre el desarrollo de gráficos de visiones múltiples.

Consideremos el concepto al que nos referiremos como D/E, el cociente entre descubrimiento y esfuerzo, un índice que resume la situación en la que manejar varios gráficos simultáneamente es tan complejo que los usuarios no son capaces de obtener ninguna conclusión de interés. Un programa de ordenador que implemente gráficos estadísticos debería intentar maximizar este índice. Debería mantener el descubrimiento alto mientras que el esfuerzo debería ser lo más bajo posible. El usuario debería de encontrarse con los gráficos tal y como se muestran arriba, sin que tenga por tanto necesidad de reorganizarlos. Optimizar D/E se convierte en crucial cuando tenemos en cuenta que los usuarios normalmente necesitan comparar una gran cantidad de modelos alternativos antes de alcanzar una conclusión y el proceso de análisis de datos es un proceso cíclico que ocurre durante periodos largos de tiempo y muchas sesiones con el paquete estadístico.
Para maximizar D/E creemos que los siguientes aspectos de gráficos estadísticos son de importancia:
Disposición: Para aumentar al máximo D/E las ventanas deben ser dispuestas automáticamente en la pantalla de una manera correcta. Los gráficos deberían aparecer dentro de lo posible en una única ventana, siendo "ventanas paneladadas" antes que ventanas separadas. El usuario debería por otro lado poder cambiar la disposición de las ventanas y mover o cambiar de tamaño estas ventanas. No obstante, debido a consideraciones de interpretabilidad es conveniente que esos cambios no distorsionen relaciones de proporcionalidad cuando no sea apropiado. Las modificaciones de la disposición deberían poder ser generales al mismo tiempo que individuales. Esto significa que, por ejemplo, cambiar de tamaño todos los gráficos simultáneamente debería ser posible mediante una acción simple.
Interacción: Para maximizar D/E, las ventanas deberían permitir que las interacciones con ellas respondan inmediatamente y que se propaguen de unas a otras de una manera apropiada. En el ejemplo del conjunto de gráficos para análisis loglinear, la selección de nuevos términos en la ventana de modelos produce automáticamente cambios en los diferentes gráficos y ventanas, de tal manera que el usuario no necesita actuar sobre ellos individualmente. Esto se consigue mediante un objeto estadístico que está especializado en el cálculo de los métodos utilizados, y en otro encargado de notificar a las demás ventanas que alguna de ellas ha cambiado y la forma en que es necesario reaccionar a ese cambio.

- **Coordinación**: Aunque sugerido en los puntos anteriores, es conveniente volver a incidir sobre la necesidad de coordinación entre gráficos estadísticos a muy diversos niveles. En [4] se intenta proporcionar un listado completo de las posibles coordinaciones entre gráficos que es posible utilizar, y [3] explica el paradigma de conexión de datos y su importancia en análisis de datos. Un ejemplo de coordinación es cuando en ciertos gráficos, la escala de los ejes debe mantenerse constante para permitir la comparabilidad [5] De este modo, cambios en la escala del eje de un gráfico deberían ser propagados a los otros gráficos para evitar que se produjeran errores de interpretación.
- **Re-análisis:** Para maximizar D/E, el modelo estadístico representado debe ocupar un papel activo en los gráficos. Si el usuario hace cambios en los gráficos que implican cambios en el modelo, las consecuencias deberían ser mostradas instantáneamente en los otros gráficos.
- **Superar las limitaciones de la pantalla:** Para maximizar D/E, debemos ser capaces de superar las limitaciones de tamaño y resolución impuestas por la pantalla. Como en el caso de los procesadores de texto (Dix, formal methods), la información que va a ser mostrada es a menudo imposible que pueda ser mostrada en la pantalla simultáneamente. En los procesadores de texto, el modelo de la ventana que se desplaza a lo largo de un rollo infinito de papel parece haberse impuesto. En el caso de los gráficos estadísticos, existe un límite práctico de alrededor ocho o nueve gráficos que pueden ser mostrados en la pantalla simultáneamente. En la actualidad, nosotros solemos programas visiones disminuidas pero potencialmente significativas (por ejemplo matrices de diagramas de dispersión) como instrumento de interacción para mostrar gráficos a tamaños más realistas (por ejemplo, interactuar sobre una celda de una de estas matrices producirá una versión ampliada de ese gráfico en un lugar prefijado). Este es un área que merece una elaboración más completa en un futuro.
- **Diversión**: Pero, por encima de todo, para maximizar D/E debemos lograr que los análisis estadísticos sean divertidos. Nuestra experiencia es que los análisis de datos implican un cierto grado de exploración basada en reglas que los usuarios (y nosotros mismos) disfrutan descubriendo. El resultado, si los conjuntos de datos utilizados lo permiten, a menudo llevan a interpretaciones que nos ayudan a entender aspectos de las cosas que por otro lado difícilmente podríamos determinar. Los gráficos dinámicos interactivos que hemos programado permiten un sistema en el que el análisis de datos es divertido. Y creemos que un analista que se está divirtiendo es uno que seguramente va a tener más descubrimientos.

Reconocimientos

Este artículo ha sido financiado en parte por el proyecto de investigación GV99-116-1-14 de la Generalitat Valencia

Referencias

0. Young, F. W. (1992). "ViSta: The Visual Statistics System". UNC Psychometric Laboratory, Chapel Hill NC.
1. Card, S. K., Mackinlay, J. D. y Shneiderman, B. (Eds.) (1999). Readings in Information Visualization :Using vision to think. Morgan Kauffman Publishers, Inc. San Francisco.
2. Wang Baldonado, M. Q., Woodruff, A. & Kuchinsky, A. (2000), "Guidelines for using Multiple Views in Information Visualization", *Proceedings of AVI 2000*, Palermo, pp. 110-119.
3. North, C. & Shneiderman, B. (2000), "Snap-Together Visualization: A User Interface for Coordinating Visualizations via Relational Schemata". In *Proceedings of AVI 2000*, Palermo.
4. Wills, G. (1999), "Linked Data Views", Statistical Computing & Statistical Graphics Newsletter, 10, 20-24.
5. Velleman, P. F. (1992), *DataDesk Handbook*, Ithaca, NY: Data Description Inc.
6. Wilhelm, A. F. X. (1999), "A data model for interactive statistical graphics". Proceedings of the Section on Statistical Graphics. Baltimore. 61-70.
7. Becker, R. A., Cleveland, W. S. and Shyu, J. (1996), "The Visual Design and Control of Trellis Display", *Journal of Computational and Statistical Graphics*, 5, 123-155.