Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

A Functional System for Statistical Graphics Presentation: brinton

Pere Millán-Martínez
UVic-UCC Servei Català de Trànsit

Ramon Oller-Piqué
UVic-UCC

Abstract

Este artículo presenta el paquete **brinton** de R desarrollado para el análisis gráfico exploratorio de datos y la selección de una representación gráfica estadística en particular. Sobre la base de **ggplot2** y **gridExtra** el paquete **brinton** introduce las funciones `wideplot()` y `longplot()` que presentan abanicos de gráficas estadísticas posibles a partir de una selección de variables. Complementaria a estas dos funciones, también introduce la función `showplot()` que permite seleccionar una gráfica específica y ajustar sus propiedades. Este conjunto de funciones se demuestra útil para entender la estructura de un conjunto rectangular de datos, dejarse sorprender por propiedades no esperadas en los datos, evaluar diferentes representaciones gráficas de éstos y seleccionar una gráfica en particular.

Keywords: Automated design, graphic design, statistical graphics, information visualization, exploratory data analysis.

1. Introduction

En 1977 R.L. Oliver (1977) vinculó la satisfacción a la expectativa en su "Expectation disconfirmation theory" (EDT). J.W. Tukey señaló que "The greatest value of a picture is when it *forces* us to notice what we never expected to see" (Tukey 1977, p.iv) lo que atribuye un valor a la gráfica relativo a la expectativa. En el campo del análisis exploratorio de datos la expectativa juega un papel especial generalmente debido a su ausencia porque las hipótesis no estan preestablecidas sinó que provienen de la observación de los datos. Las hipótesis o la definición del problema, como señaló J. Bertin ese mismo preciso año (Bertin 1977, p.2), no son automatizables y aquí nace el reto de presentar gráficas a un usuario para que éste observe los datos, pueda plantear una hipótesis y afinar la selección de la gráfica estadística que dé respuesta a estas hipótesis.

Puestos a elegir una estrategia para presentar gráficas a un usuario, Millán-Martínez y Valero-Mora (2018), por ejemplo, diferencian las estrategias según si éstas se basan en las características de los datos – *functional design*– (Benson and Kitous 1977, Gnanamgari (1981), Kamps (1999), Valero-Mora, Ledesma, and Friendly (2012)), en los hábitos de un usuario – *content-based filtering* –, en los hábitos de un grup de usuarios – *collaborative filtering* – (Mutlu, Veas, Trattner, and Sabol 2015), en las tareas a realizar por los usuarios – *task design* – (Bowman 1967, Casner (1991)), en las características de la percepción humana – *perceptual design* – (Cleveland and McGill 1984, Mackinlay (1986)), en las limitaciones del canal de comunicación o la pantalla en la que se proyectan las gráficas – *responsive design* – (Gnanamgari 1981) y, finalmente, en la selección de carácterísticas de la gráfica deseada o modelos de representación – *this could be called model design* – (Roth, Kolojejchick, Mattis, and Goldstein 1994). Si, como hemos dicho anteriormente, carecemos de hipótesis, si obviamos las limitaciones del hardware y el recuerdo en la selección de gráficas realizado en otras ocasiones, queda la posibilidad de acotar las gráficas a presentar según las características de los datos, las características de la percepción humana y la selección de modelos de representación.

En el entorno de programación R, hay implementadas diferentes de las estrategias antes citadas. El *functional design* se encuentra, por ejemplo, implementada en la función `plot()` que, si la aplicamos al dataset `cars` produce un diagrama de dispersión (*scatterplot*) dado que éste contiene dos variables numéricas y es de classe `data.frame` mientras que, si lo aplicamos al dataset `airmiles` produce un diagrama de línea (*line graph*) porque éste contiene una única variable numérica y es de clase `ts`. El *task design* se encuentra implementado en multitud de librerías como por ejemplo **survminer** (Therneau 2015) que incluye la función `ggsurvplot()` para componer gráficas específicas para el análisis de supervivencia (*survival analysis*). El *model design* lo encontramos en funciones básicas como por ejemplo `barplot()` que produce un diagrama de barras (*bar graph*), `hist()` que produce un histograma(*histogram*) o `pie()` que produce un diagramas de pastel (*pie charts*). El *perceptual design* también se encuentra implementado en librerías como por ejemplo **ggplot2** (Wickham 2016) en aspectos como el tamaño, forma o el color de los puntos que asigna por defecto, las líneas de ayuda (*grid lines*) o el color del fondo del panel (*panel background color*).

Este artículo presenta el paquete **brinton** que explora una nueva estrategia implementada sobre en el entorno de programación R y las librerías **ggplot2** y **gridExtra** entre otras. Esta estrategia consiste en presentar al usuario un abanico de gráficas posibles a partir de las características de los datos para que, una vez observados, el usuario pueda plantearse preguntas y explorar nuevos catálogos de gráficas o una gráfica en particular.

2. Multipanel graphics

Existen diferentes tipos de gráficas multipanel según la variadad de gráficas y el origen de los datos que éstas muestran. Por un lado tenemos los cuadros de mando (*dashboards*) que generalmente combinan en un espacio limitado, diferentes tipos de gráficas en diferentes paneles, cada una de las cuales puede representar datos de diferentes orígines y cuya utilidad es monitorizar procesos complejos. Por otro lado tenemos las gráficas condicionadas (*conditioning plots*)[1] que muestran un mismo tipo de gráfica que se repite en diferentes paneles que

[1]La terminología de las gráficas condicionadas no és unánime. Primero fueron primero descritas por J. Bertin como *séries homogènes* (Bertin 1967, p.?), luego E.Tufte las dio a conocer como *small multiples* (Tufte

generalmente conservan la misma escala y que representan diferentes subconjuntos de unos datos según una o dos variables. Otro ejemplo de gráficas multipanel lo tenemos las matrices de gráficas (*matrix of plots*) que relacionan pares de variables de un conjunto de datos, como por ejemplo las matrices de diagramas de dispersión (*scatterplot matrix*) (Hartigan 1975) o las matrices de gráficas de elipse HE (*HE plots*) (Friendly 2007).

Existen básicamente tres vías para graphicar datos en R. La primera es utilizar *base graphics* y, si se pretende componer gráficas multipanel

y por consiguiente cuesta solicitar a un sistema la presentación de una gráfica u otra. Esto genera la conveniencia de utilizar métodos gráficos que, por un lado, relacionen los valores de las diferentes variables y, por otro lado, que estos métodos gráficos incluyan una variedad de tipos de gráficas que permitan responder a diferentes preguntas que se puedan formular.

2.1. Code formatting

Don't use markdown, instead use the more precise latex commands:

- R
- **brinton**
- `print("abc")`

3. R code

Can be inserted in regular R markdown blocks.

```
R> x <- 1:10
R> x

 [1]  1  2  3  4  5  6  7  8  9 10
```

References

Becker RA, Cleveland WS, Shyu MJ (1996). "The visual design and control of trellis display." *Journal of computational and Graphical Statistics*, **5**(2), 123–155.

Benson WH, Kitous B (1977). "Interactive analysis and display of tabular data." ACM.

Bertin J (1967). *Sémiologie graphique. Les diagrammes, les réseaux, les cartes*. Mouton, Paris.

Bertin J (1977). *La graphique et le traitement graphique de l'information*. Flammarion, Paris.

1983), W.S. Cleveland se refirió a ellas como *juxtaposed panels* (Cleveland 1985, p.200) y también como *trellis graphics* (Becker, Cleveland, and Shyu 1996), en el entorno de R de conocen básicamente como *lattice graphics* (Sarkar 2008).

Bowman W (1967). *Graphic communication*. Wiley series on human communication. Wiley.

Casner SM (1991). "Task-analytic approach to the automated design of graphic presentations." *ACM Transactions on Graphics (TOG)*, **10**(2), 111–151.

Cleveland W (1985). *The Elements of Graphing Data*. Hobart Press, Summit, New Jersey.

Cleveland WS, McGill R (1984). "Graphical perception: Theory, experimentation, and application to the development of graphical methods." *Journal of the American statistical association*, **79**(387), 531–554.

Friendly M (2007). "HE plots for Multivariate General Linear Models." *Journal of Computational and Graphical Statistics*, **16**(4), 421–444.

Gnanamgari S (1981). *Information presentation through default displays*. University of Pennsylvania. Ph.D. dissertation.

Hartigan JA (1975). "Printer graphics for clustering." *Journal of Statistical Computation and Simulation*, **4**(3), 187–213.

Kamps T (1999). *Diagram Design: A Constructive Theory*. Springer Berlin Heidelberg.

Mackinlay J (1986). "Automating the design of graphical presentations of relational information." *ACM Trans. Graph.*, **5**(2), 110–141.

Millán-Martínez P, Valero-Mora P (2018). "Automating statistical diagrammatic representations with data characterization." *Information Visualization*, **17**(4), 316–334.

Mutlu B, Veas E, Trattner C, Sabol V (2015). "VizRec: A Two-Stage Recommender System for Personalized Visualizations." In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, IUI Companion '15, pp. 49–52. ACM, New York, NY, USA.

Oliver RL (1977). "Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation." *Journal of applied psychology*, **62**(4), 480.

Roth SF, Kolojejchick J, Mattis J, Goldstein J (1994). "Interactive graphic design using automatic presentation knowledge." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 112–117. ACM.

Sarkar D (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5, URL http://lmdvr.r-forge.r-project.org.

Therneau TM (2015). *A Package for Survival Analysis in S*. Version 2.38, URL https://CRAN.R-project.org/package=survival.

Tufte ER (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire.

Tukey J (1977). *Exploratory Data Analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company. ISBN 9780201076165. URL https://books.google.es/books?id=UT9dAAAAIAAJ.

Valero-Mora P, Ledesma RD, Friendly M (2012). "The History of ViSta: The Visual Statistics System." *Wiley Interdisciplinary Reviews: Computational Statistics*, **4**(3), 295–306.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Affiliation:

Pere Millán-Martínez
UVic-UCC Servei Català de Trànsit
Carrer Diputació, 355 08009 Barcelona
E-mail: info@sciencegraph.org
URL: http://sciencegraph.org

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd