Análisis y Tratamiento de Datos con R

Con ejemplos e ilustraciones

Primera Edición

Diego Paul Huaraca S.
MS-PLUS, Inc.

MS+
Soluciones
Matemáticas

Un aporte de Source Stat Lab Ecuador a la sociedad.

Índice general

1

Introducción

Hace varios años en el área de la Matemática y la Estadística existía un claro concenso acerca del lenguaje de programación que se debía enseñar en las aulas universitarias, este era `Pascal`. Pues se le consideraba un lenguaje sencillo y al mismo tiempo elegante, en la actualidad son pocas las personas que consideran que Pascal sea adecuado debido a sus deficiencias como: la ausencia de modularidad, la falta de paradigmas de programación (OOP[1]), el desarrollo de software libre, etc.

En la Estadística, aprender a programar es un paso importante para acercarse a la comprensión de la información. Por esta razón, creo firmemente que la programación es una habilidad vital para todos los que trabajamos analizando datos.

En Ecuador, el `lenguaje de programación R` aún no goza de un amplio reconocimiento en las aulas universitarias, sin embargo, mi interés me ha llevado a usar el programa desde el año 2010 por tres propiedades esenciales:

- **Reproducibilidad:** Capacidad para recrear un análisis pasado.
- **Automatización:** Capacidad para volver a crear un análisis cuando han surgido cambios en los datos.
- **Comunicación:** El código es solo texto, por lo que es fácil de comunicar y compartir. Esto hace simple conseguir ayuda de usuarios de todo el mundo.

Por esta razón, a través de este texto deseo compartir todo el conocimiento adquirido en esta larga pero al mismo tiempo entretenida travesía con el lenguaje R.

1.1. Lenguaje R

Para introducirnos en el *mundo del lenguaje R* debemos tener claro la doble naturaleza del programa como: *entorno* y *lenguaje de programación*, pues el mismo integra un conjunto de herramientas y comandos empleados en el tratamiento de datos, la realización de cálculos, el análisis estadístico, así como la representación gráfica en alta calidad y la reportería dinámica.

[1] Programación orientada a objetos

Figura 1.1: Logo del proyecto R

El lenguaje R se encuentra enmarcado como un software estadístico flexible, potente y profesional que se distribuye *libremente*[2] bajo licencia GNU (General Public License), y en la actualidad es muy utilizado por la comunidad científica.

Las libertades esenciales que la licencia GNU permite al usuario son:

- La libertad de ejecutar el programa como se desea, con cualquier propósito;
- La libertad de estudiar cómo funciona el programa, y cambiarlo para que haga lo que usted quiera. El acceso al código fuente es una condición necesaria para ello.;
- La libertad de redistribuir copias para ayudar a su prójimo;
- La libertad de distribuir copias de sus versiones modificadas a terceros. Esto le permite ofrecer a toda la comunidad la oportunidad de beneficiarse de las modificaciones.

Dadas sus características el lenguaje R tiene gran potencial para ser utilizado en diferentes áreas de la estadística, finanzas, simulación, reportería dinámica, biomatemática, minería de datos, big data, etc., y puede ser instalado en diversas plataformas y sistemas operativos tales como: Windows, Linux, Mac OS X y Unix.

1.2. Historia

R inició como un proyecto experimental para utilizar métodos de Lisp[3] en la construcción de un pequeño banco de pruebas que sirva para evaluar posibles construcciones de entornos estadísticos. Desde el inicio del proyecto se decidió usar la sintaxis del lenguaje S[4]. Como consecuencia, la sintaxis del lenguaje R es similar al lenguaje S, pero la semántica que aparentemente vemos es parecida a la de S, en realidad es sensiblemente diferente, sobre todo en los detalles un poco más profundos de la programación.

Ross Ihaka inicia el proyecto R tras haber obtenido acceso a cierta información importante sobre el lenguaje S, la cual fue publicada por John Chambers y Rick Becker, creadores del lenguaje S. En corto tiempo, Ross nota las similitudes existentes entre S y Scheme[5], en su afán de mostrarse ante Alan Zaslavsky[6] le propone indicar el uso del ámbito léxico para la obtención de variables propias. Sin ninguna copia de Scheme trata de mostrarle usando S, sin embargo sus intentos fracasan debido a las diferencias entre las reglas de asignación que tiene S y Scheme, esto le llevó a darse cuenta que S requería de ciertas funcionalidades extras con la finalidad de lograr

[2] *Software libre* es el software que respeta la libertad de los usuarios y la comunidad. A grandes rasgos, significa que los usuarios tienen la libertad de ejecutar, copiar, distribuir, estudiar, modificar y mejorar el software.

[3] Lenguaje de programación multiparadigma creado en 1958 por el MIT.

[4] Lenguaje de programación estadístico comercial desarrollado en Bell Laboratories.

[5] Lenguaje de programación desarrollado en 1970 por el MIT.

[6] Alan M. Zaslavsky, PhD, profesor de Estadística del Departamento de Cuidado de la Salud en la Escuela Médica de Harvard.

convertirle en un completo entorno estadístico.

Tiempo más tarde, Ross Ihaka y Robert Gentleman llegan a ser colegas en la Universidad de Auckland[7], y ambos muestran interés en temas de Estadística Computacional. Como profesores del Departamento de Estadística vieron la necesidad de mejorar un ambiente del laboratorio de computación e inician su trabajo con la visión de crear un lenguaje similar al S pero con más funcionalidades, mismas que ya fueron identificadas previamente por Ross en su intento de incluir los ámbitos léxicos.

Figura 1.2: Ross Ihaka & Robert Gentleman

El desarrollo del lenguaje R como tal inició en el año de 1991, el progreso fue bastante bueno por lo que para Agosto de 1993 decidieron colocar algunas copias binarias del lenguaje R en Statlib[8] y a su vez anunciaron al público el lanzamiento de la versión alfa del programa por medio de la lista de noticias de S (*S news mailing list*). De manera sorpresiva un gran número de personas probaron el nuevo lenguaje y ofrecieron su retroalimentación sobre la versión que habían liberado, entre ellos el más persistente fue Martin Machler[9] que los animó a liberar el código fuente de manera que R se distribuya bajo licencia GNU General Public License, por lo cual en Junio de 1995 aparece la primera versión libre (*open source*), el interés por el lenguaje R creció rapidamente al punto que para Marzo de 1996 fue necesario crear la propia lista de noticias y un año más tarde se tuvo que reemplazar por listas específicas como: R-announce, R-help y R-devel, esto debido a la gran cantidad de consultas que realizaban los usuarios en varios temas relacionados a la instalación, funcionamiento, etc.

R es considerado la versión libre del programa comercial S-Plus, el cual fue desarrollado para AT&T Bell Laboratories por John M. Chambers y colaboradores en el año 1988, aunque son evidentes las diferencias entre R y S, la gran mayoría del código escrito para S funciona sin inconvenientes en R. En la actualidad el lenguaje S es distribuido por la empresa TIBCO bajo el nombre de S-PLUS.

1.3. R Core Group

Para a mediados de 1997 se estableció el `R Core Group` o `R Core Team`, un grupo de desarrolladores talentosos y experimentados con permisos para administar el código fuente del lenguaje R, en sus inicios lo conformaban los siguientes miembros:

[7]Universidad pública situada en Auckland. Fue fundada en 1883 como parte de la Universidad de Nueva Zelanda.

[8]Sistema para la distribución de software estadístico vía electrónica.

[9]Profesor de Estadística Computacional, Departamento de Matemática, Escuela Politécnica Federal de Zúrich.

- Ross Ihaka
- Robert Gentleman
- Martin Machler
- Doug Bates
- Peter Dalgaard
- Kurt Hornik
- Friedrich Leisch
- Thomas Lumley
- Paul Murrell
- Heiner Schwarte
- Luke Tierney

Entre sus tareas principales se encontraba realizar los cambios en el código fuente de acuerdo al informe de errores (*bugs*) reportados por los usuarios, además aportaban sustancialmente sugiriendo varias mejoras, vale mencionar que todas las tareas se realizaban de manera voluntaria dentro del grupo, en la actualidad lo conforman alrededor de 20 personas que se encuentran en 11 diferentes cuidades del mundo.

Con la participación de los miembros del R Core Group se fundó además la `R Foundation for Statistical Computing`, una organización sin fines de lucro con oficinas Viena que en la actualidad se encuentra a cargo de la distribución del software.

1.4. CRAN

En la actualidad el software se distribuye gratuitamente a través del repositorio `Comprehensive R Archive Network` (CRAN) propiedad de la `R Foundation for Statistical Computing` por medio del enlace `http://www.r-project.org`, su mantenimiento se encuentra a cargo del grupo R Core Team desde 1997 asistido por una gran cantidad de colaboradores internacionales.

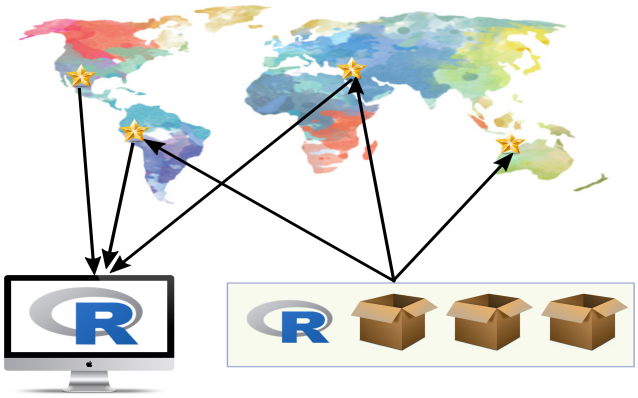
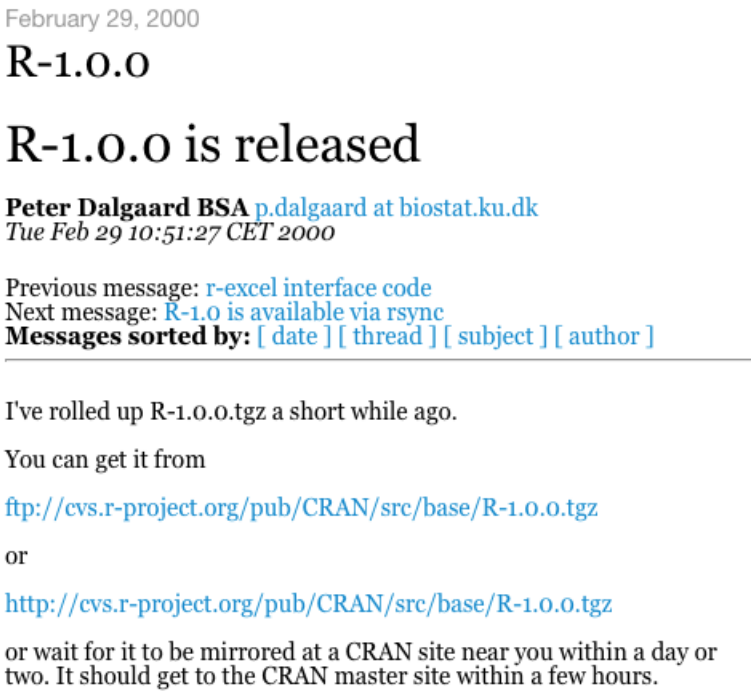


Figura 1.3: Funcionamiento CRAN

El repositorio CRAN es fundamental para el uso del lenguaje R, pues en el sitio se almacena el ejecutable del programa así como las librerías que permiten ampliar sus capacidades. Con la finalidad de evitar el colapso del *mundo estadístico* no se tiene una única ubicación al que todo mundo tiene acceso, el CRAN se *refleja* en diferentes lugares de todo el mundo, de esta manera como residente de Ecuador podría acceder a una ubicación CRAN en Ecuador[10], mientras que si se encuentra en un país que no tiene repositorio lo más recomendable es acceder a la copia del repositorio de un país cercano. Los países desarrollados como: EEUU, Alemania, Francia tienen múltiples CRAN's para abastecer a los usuarios. *La filosofía básica es elegir un repositorio que*

[10]En Ecuador se cuenta con el CRAN de la Escuela Politécnica del Litoral

se encuentre geográficamente cercano al usuario.



February 29, 2000

R-1.0.0

R-1.0.0 is released

Peter Dalgaard BSA p.dalgaard at biostat.ku.dk
Tue Feb 29 10:51:27 CET 2000

Previous message: r-excel interface code
Next message: R-1.0 is available via rsync
Messages sorted by: [date] [thread] [subject] [author]

I've rolled up R-1.0.0.tgz a short while ago.

You can get it from

ftp://cvs.r-project.org/pub/CRAN/src/base/R-1.0.0.tgz

or

http://cvs.r-project.org/pub/CRAN/src/base/R-1.0.0.tgz

or wait for it to be mirrored at a CRAN site near you within a day or two. It should get to the CRAN master site within a few hours.

Figura 1.4: Lanzamiento versión 1.0.0

Se tuvo que esperar hasta el 29 de Febrero del 2000, para que sus desarrolladores consideren que el software se encontraba completo y estable al punto que liberaron la versión 1.0.0. En el siguiente enlace se puede encontrar cronológicamente todos los cambios que se han ido realizando sobre el lenguaje R `http://timelyportfolio.github.io/rCharts_timeline_r/`.

1.5. Soporte

El lenguaje R al ser un software libre carece de soporte técnico, sin embargo, debido a la gran cantidad de usuarios que ha adquirido en los últimos años han aparecido empresas que proveen varios tipos de soportes bajo pago. Entre las más destacadas se encuentran:

- RStudio, Inc.
- Revolution Analytics, Inc.
- XL - Solutions Corporation.
- Quantide, Inc.
- R-Plus
- Statconn

De entre las empresas antes mencionadas se destaca de sobremanera Revolution Analytics[11], Inc. la cual se tomó la tarea de optimizar el código fuente del lenguaje R y ha permitido que el programa sea multicore, es decir, a través de estas modificaciones se logró aprovechar al máximo la funcionalidad de todos los núcleos de los procesadores (16 núcleos para procesadores Core i7, etc.). El impacto inmediato del multicore se ve reflejado en la mayor velocidad de procesamiento de información permitiendo de este modo trabajar con grandes cantidades de datos sin ningún problema, así como también en la reducción del tiempo de ejecución.

[11]Revolution Analytics fue fundada en el año 2007 con la finalidad de dar soporte comercial al software Revolution R, adicionalmente provee componentes tales como: ParallelR, RevoScaleR, RevoDeployR, etc.

Figura 1.5: Revolution R Open

El programa optimizado es 100 % compatible con el software relacionado a R (paquetes, interfaces, etc.) y puede obtenerse desde el sitio web `http://www.revolutionanalytics.com` como `Revolution R Open` de forma gratuita bajo licencia GPL, sin embargo en el caso de requerir soporte se puede adquirir la versión pagada `Revolution R Enterprise`.

1.6. Funcionamiento

El programa R es un lenguaje orientado a objetos (OOP[12]) diseñado en un entorno auténtico bajo el cual esconde su simplicidad y flexibilidad, lo cual permite a sus usuarios añadir funcionalidad mediante la definición de nuevas *funciones*. El término *orientado a objetos* hace referencia a un paradigma de la programación que emplea objetos en sus interacciones y diseño de aplicaciones. R almacena sus variables, datos, funciones, resultados, etc., en la memoria activa del computador en forma de objetos con un nombre específico y pueden ser modificados o manipulados por el usuario mediante operadores y funciones.

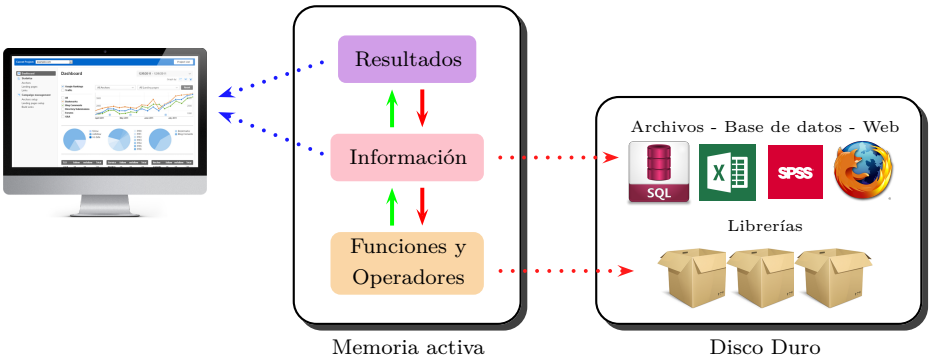


Figura 1.6: Esquema del funcionamiento de R

El hecho que R sea un lenguaje de programación puede desmotivar a muchos usuarios, los cuales piensan que para iniciarse en el programa se necesita *alma de programador* lo cual no es cierto. Primero R es un lenguaje interpretado similar a Java, y segundo no es un lenguaje compilado a diferencia de C, C++, Fortran, Pascal, ect. sino más bien mediante comandos ingresados por teclado los cuales se ejecutan directamente sin necesidad de construir archivos ejecutables. En su mayoría R se encuentra programado en C++, Python y Fortran, esto implica que R tiene la capacidad de interpretar código externo mediante el uso de ciertas librerías. La utilidad básica de lo anterior se encuentra en elaborar scripts, por ejemplo en C++, que emplean menor tiempo de ejecución.

[12]La programación OOP está basada en varias técnicas: herencia, clasificación, identidad, polimorfismo y encapsulamiento.

El programa R incluye 8 bibliotecas o paquetes estándar, sin embargo, las capacidades de R pueden ser ampliadas fácilmente mediante la incorporación de paquetes que se encuentran disponibles en varios repositorios como:

- CRAN
- BioConductor
- Github
- Omegahat
- MRAN
- RForge, entre otros.

Los paquetes estándar pueden ser visualizados a través del comando:

```
search()

##  [1] ".GlobalEnv"         "package:data.table" "package:dplyr"
##  [4] "package:knitr"      "package:stats"      "package:graphics"
##  [7] "package:grDevices"  "package:utils"      "package:datasets"
## [10] "package:methods"    "Autoloads"          "package:base"
```

En la actualidad[13] existen 7190 paquetes válidos en el repositorio CRAN, que se encuentran ordenados por fecha de publicación o alfabéticamente agrupadas en diversas líneas de investigación.

1.7. Ventajas

Entre las principales ventajas que posee el software R podemos anotar lo siguiente:

- Al tratarse de un software libre el costo es nulo.
- Se han implementado una gran cantidad de métodos estadísticos desde los más básicos hasta los más avanzados y modernos. Todos los métodos se encuentran organizados en librerías, que se encuentran en constante crecimiento.
- Capacidad para acceder a datos de múltiples formatos. En la actualidad existen varias librerías para leer datos desde SPSS, SAS, STATA, MySQL, Excel, etc.
- Gran capacidad para la manipulación de datos y funciones, así como para la generación de gráficos de alta calidad; estos últimos pueden ser almacenados en varios formatos dependiendo del sistema operativo (jpg, png, bmp, ps, pdf, emf, gif, xfig, etc.).
- Facilidad para enlazarse con LaTeX y generar reportes dinámicos.
- Amplia bibliografía tanto en internet como en libros publicados por prestigiosas editoriales como: Springer, Wiley, O'Reilly, Chapman & Hall/CRC, etc.
- Fácil visualización e interpretación de los algoritmos implementados en R con lo cual el usuario puede conocer exactamente lo que el ordenador ejecuta.
- Permite visualizar los algoritmos en él implementados, modificarlos y ajustarlos a nuestras necesidades (esto no es permitido con el software comercial).
- Una gran comunidad de usuarios, lo cual permite obtener ayuda de manera fácil de expertos por medio de foros tales como: StackOverFlow, R mailing lists, entre otros.

[13] Información obtenida al 21 de Septiembre 2015.

1.8. Desventajas

Los inconvenientes a los cuales se deben enfrentar los usuarios de R son:

- Al ser un programa libre carece de un departamento de atención al cliente al cual se pueda recurrir en caso de que se reporte un inconveniente con el mismo. Sin embargo, existe una comunidad en crecimiento de usuarios de R que se encuentran dispuestos a colaborar desinteresadamente en la resolución de problemas.
- El software R como tal no dispone de una interfaz amigable para el usuario, las tareas se llevan a cabo a través de líneas de comando lo cual puede resultar difícil para el usuario común. No obstante con el desarrollo de IDE's se ha facilitado en gran medida la experiencia del programa con el usuario común.
- El código en R es interpretado, no compilado, lo cual puede ocasionar una ejecución lenta en ocasiones en las que se realizan simulaciones intensas. Con el fin de remediar lo anterior el grupo R Core Team a partir de la versión 2.14 ha precompilado todas las funciones y librerías de R con el objetivo de acelerar la ejecución.
- R no es particularmente un lenguaje de programación rápido, si a eso sumamos que muchos usuarios escriben pobremente su código, obtenemos como resultado un funcionamiento lento.

1.9. Usuarios en el mundo

Han transcurrido 23 años desde el aparecimiento de R y el número de usuarios sigue creciendo día a día[14]. Lo que comenzó como un proyecto para proporcionar un *software estadístico*, hoy en día se ha convertido en una solución estándar para el análisis de datos en todo el mundo.

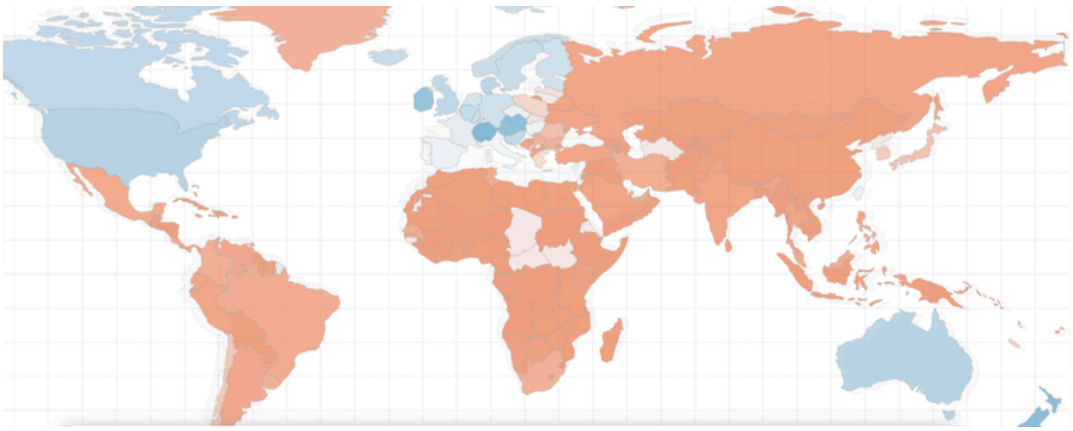


Figura 1.7: Actividad de los usuarios

Los países con mayor cantidad de usuarios activos son:

1. Suiza
2. Nueva Zelanda
3. Austria
4. Irlanda
5. Estados Unidos
6. Australia
7. Singapur
8. Dinamarca

[14]Se estima que existe alrededor de dos millones de usuarios en todo el mundo.

9. Reino Unido
10. Canadá

Llama la atención que cada vez más se incrementa el número de cursos y cantidad de libros relacionados con el programa R. Además los usuarios optan por escribir en sus blogs varias de sus experiencias con ciertos paquetes lo cual se ha convertido en material de gran ayuda para los principiantes.

1.9.1. UseR!

Con la finalidad de conocer a otros usuarios, aprender de los expertos, compartir experiencias y participar de tutoriales se ha venido llevando a cabo conferencias internacionales anualmente desde el 2004.

Figura 1.8: Logo UseR!

La última conferencia `UseR!` se llevó a cabo entre el 30 de Junio y 3 de Julio de 2015 en la ciudad de Aalborg, Dinamarca y contó con más de 660 participantes de 40 diferentes países. Mientras que para el año 2016, la conferencia tendrá lugar en la Universidad de Stanford (California, EEUU) entre el 27 y 30 de Junio.

Para mayor información: `https://www.r-project.org/conferences.html`

1.9.2. R User Groups

Debido que la comunidad de usuarios a nivel mundial es demasiado grande, existen grupos más pequeños formados por: temas especificos, idioma, ubicación, etc., que permiten interactuar de manera sencilla con otros usuarios. Actualmente, existen alrededor de 150 grupos registrados en el directorio de Revolution Analytics[15], la gran mayoría de ellos se reúnen periodicamente y cuentan con una página web en Meetup que les permite compartir noticias y detalles de próximas reuniones.

1.9.3. Usuarios de R Ecuador

Con la idea de iniciar un espacio que permita compartir conocimientos, experiencias e inquietudes sobre el lenguaje R, el 26 de Septiembre del año 2015 se fundó el grupo de `Usuarios de R Ecuador` en las instalaciones de la Escuela Politécnica Nacional (Quito).

El grupo cuenta con los siguientes enlaces abiertos para fututos nuevos usuarios formen parte del grupo:

- **Meetup**: `http://www.meetup.com/es-ES/Usuarios-de-R-Ecuador/`
- **Github**: `https://github.com/UsuariosREc`
- **Website**: `http://usuariosrec.github.io/Web/`

[15] `http://blog.revolutionanalytics.com/local-r-groups.html`