

---

**LIBRO DE RESÚMENES**  
**V JORNADAS DE USUARIOS DE R**  
**CENTRO DE ARTE Y TECNOLOGÍA ETOPIA, ZARAGOZA**  
**12 Y 13 DE DICIEMBRE DE 2013**

---

COMITÉS ORGANIZADOR Y CIENTÍFICO  
[HTTP://R-ES.ORG/5J](http://R-ES.ORG/5J)

1 DE DICIEMBRE DE 2013



# V Jornadas de Usuarios de



Centro de Arte y Tecnología, Etopía  
Zaragoza, 12 y 13 de Diciembre

Inscripción: <http://r-es.org/5j>

Premio **TELEFÓNICA DIGITAL**  
a la mejor aplicación R en Big Data

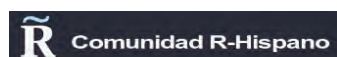


Premio **SYNERGIC PARTNERS**  
II Concurso de análisis de datos con R

**PATROCINAN:**



**Universidad  
Zaragoza**



---




© 2013 Organización de las V Jornadas de Usuarios de R



Esta obra está bajo una licencia **Reconocimiento-No comercial-Compartir bajo la misma licencia** 3.0 España de Creative Commons. Para ver una copia de esta licencia, visite:

<http://creativecommons.org/licenses/by-nc-sa/3.0/es/legalcode.es>.

Usted es libre de copiar, distribuir y comunicar públicamente la obra, y hacer obras derivadas bajo las condiciones siguientes:

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra. Alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor. Nada en esta licencia menoscaba o restringe los derechos morales del autor.

# Índice general

---

Índice general	III
<b>Información General</b>	<b>VII</b>
Presentación	VIII
Información útil	X
Comité organizador	XII
Comité científico	XIII
Patrocinadores	XIV
Programa	XV
<b>I Sesión de Comunicaciones Jueves</b>	<b>1</b>
<b>1 Evaluación del uso de modelos mixtos para estimación de la tasa de paro con poca muestra</b>	<b>2</b>
<i>José Luis Cañadas Reche</i> <i>Instituto de Estudios Sociales Avanzados (IESA-CSIC)</i>	
<b>2 Algunos aspectos prácticos del manejo de datos de encuesta con R</b>	<b>3</b>
<i>Jesús Bouso Freijo</i> <i>Centro de Investigaciones Sociológicas (CIS)</i>	
<b>3 Package xkcd: Plotting ggplot2 graphics in a XKCD style</b>	<b>4</b>
<i>Emilio Torres Manzanera</i> <i>Universidad de Oviedo</i>	
<b>4 Métrica de Wasserstein para la comparación de matrices origen-destino</b>	<b>5</b>
<i>Aleix Ruiz de Villa, Jordi Casas, Martijn Breen</i> <i>TSS - Transport Simulation Systems</i> <i>RugBcn, Grupo de usuarios de Barcelona</i>	

<b>5 Mejora de la detección visual de datos atípicos mediante una modificación en las caras de Chernoff</b>	<b>6</b>
<i>Beatriz González Pérez, Victoria López López, Jorge Cordero</i> <i>Universidad Complutense de Madrid</i>	
<b>6 Categorización automática de contenidos web con R</b>	<b>7</b>
<i>Pedro Concejero, César García, Ana Armenta, Paulo Villegas, J. Gregorio Escalada</i> <i>Telefónica Digital, Product Development and Innovation</i>	
<b>7 Sesgo de publicación en ciencias médicas</b>	<b>11</b>
<i>Borja Santos Zorrozuía, Eduardo González Fraile, Javier Ballesteros Rodríguez</i> <i>Universidad del País Vasco (UPV/EHU)</i> <i>Cibersam</i> <i>Programa PREDOC Gobierno Vasco</i> <i>Instituto de Investigaciones Psiquiátricas</i>	
<b>8 Tratamiento de datos con R para control de calidad basado en valoraciones biológicas. Rectas Paralelas.</b>	<b>13</b>
<i>Faustino Huertas Muñoz, María Victoria Collazo López, Gloria Frutos Cabanillas</i> <i>Agencia Española de Medicamentos y Productos Sanitarios (AEMPS)</i> <i>Dpto. de Estadística e Investigación Operativa. Facultad de Farmacia. UCM</i>	
<b>9 Análisis exploratorio de datos del mercado eléctrico español con R</b>	<b>15</b>
<i>J.M. Velasco, B. González, G. Miñana, R. Caro, H. Marrao, J. Gil, V. López</i> <i>Departamento de Arquitectura de computadores y automática. Universidad Complutense de Madrid.</i> <i>Indizen Technologies, S.L.</i>	
<b>10 Utilidad clínica de modelos predictivos: análisis mediante funciones de densidad de probabilidad estimadas por métodos tipo kernel</b>	<b>16</b>
<i>Luis Mariano Esteban, Gerardo Sanz, Ángel Borque, José Rubio Briones</i> <i>Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza</i> <i>Departamento de Métodos estadísticos, Universidad de Zaragoza</i> <i>Departamento de Urología. Hospital Universitario Miguel Servet, Zaragoza</i> <i>Departamento de Urología. Instituto Valenciano de Oncología, Valencia</i>	
<b>II Sesión de Comunicaciones Viernes</b>	<b>18</b>
<b>11 El paquete W2CWM2C: análisis de correlación de wavelet. Casos bivariado y multivariado.</b>	<b>19</b>
<i>Josué M. Polanco Martínez</i> <i>Instituto de Economía Pública y Dept. de Econometría y Estadística, Universidad del País Vasco</i>	
<b>12 Desarrollo de Interfaces Web utilizando programación funcional en R</b>	<b>21</b>
<i>Jorge Luis Ojeda Cabrera</i> <i>Dept- Métodos Estadísticos, Univ. de Zaragoza</i>	

<b>13 Análisis Automatizado de Cuasi-Implicaciones el Proyecto RCHIC: primeros pasos</b>	<b>22</b>
<i>Rubén Pazmiño, Raphael Couturier, Pablo Gregori</i> <i>Escuela Superior Politécnica de Chimborazo. Ecuador</i> <i>Universida Comte. Francia</i> <i>Universidad Jaume I. España</i>	
<b>14 Postprocesado de resultados de análisis de elementos finitos con R</b>	<b>24</b>
<i>Andres Sanz Garcia, Julio Fernandez Cenicerros, Rubén Urraca Valle, Roberto Fernandez Martinez</i> <i>Division of Bioscience. University of Helsinki, Finland</i> <i>EDMANS. Universidad de La Rioja, Spain</i> <i>TELEVITIS. Universidad de La Rioja, Spain</i> <i>Department of Mining and Metallurgical Engineering and Materials Science. University of Basque Country, Spain</i>	
<b>15 Preprocesado de imágenes hiperespectrales en R</b>	<b>25</b>
<i>Rubén Urraca Valle, Borja Millán, Roberto Fernandez Martinez, Andres Sanz Garcia</i> <i>TELEVITIS. Universidad de La Rioja, Spain</i> <i>Department of Mining and Metallurgical Engineering and Materials Science. University of Basque Country, Spain</i> <i>Division of Bioscience. University of Helsinki, Finland</i>	
<b>16 Análisis clasificatorio de la actividad electroencefalográfica a través del paso de señales temporales al dominio de la frecuencia</b>	<b>27</b>
<i>Roberto Fernandez Martinez, Rubén Lostado Lorza, Rubén Urraca Valle, Andres Sanz Garcia</i> <i>Department of Mining and Metallurgical Engineering and Materials Science. University of Basque Country, Spain</i> <i>Universidad de La Rioja. Spain</i> <i>Universidad de La Rioja. Spain</i> <i>Division of Bioscience. University of Helsinki, Finland</i>	
<b>17 Medición de la potencia en deportistas usando R y encoders</b>	<b>28</b>
<i>Xavier de Blas Foix</i> <i>Universitat Ramon Llull</i> <i>FPCEE Blanquerna</i> <i>Grupo SAFE</i> <i>Chronojump-Boscosystem</i>	
<b>18 Estrategias de Captación de Clientes en Mercados con Competencia</b>	<b>29</b>
<i>Francisco Jesús Rodríguez Aragón</i> <i>Associate Professional Risk Manager</i>	
<b>19 Previsión de equipamientos educativos, culturales y sanitarios en los barrios de nueva creación de la ciudad de Zaragoza</b>	<b>31</b>
<i>Sergio Jiménez Sanjuán</i> <i>SCIEN Analytics</i>	

<b>20 Simulación en R de modelos definidos en hoja de cálculo</b>	<b>33</b>
<i>Ramiro Serrano García, Gregorio R. Serrano</i>	
<i>Keller Graduate School of Management</i>	
<i>Universidad Complutense de Madrid</i>	
<b>III Talleres</b>	<b>34</b>
<b>21 Relenium, selenium en R. Un nuevo paquete para webscraping.</b>	<b>35</b>
<i>Aleix Ruiz de Villa, Lluís Ramon, Andreu Vall</i>	
<i>TSS - Transport Simulation Systems</i>	
<i>RugBcn, Grupo de usuarios de Barcelona</i>	
<b>22 Big data analytics: R + Hadoop</b>	<b>36</b>
<i>Carlos J. Gil Bellosta</i>	
<i>Datanalytics</i>	
<b>23 Cazando información espectro-temporal en datos ambientales con R</b>	<b>37</b>
<i>Josué M. Polanco Martínez</i>	
<i>Instituto de Economía Pública y Dept. de Econometría y Estadística, Universidad del País Vasco</i>	
<b>24 Docencia de R mediante investigación reproducible. 'RStudio', 'knitr', 'mark-down'</b>	<b>40</b>
<i>Jose Antonio Palazon Ferrando, Antonio Maurandi López</i>	
<i>Universidad de Murcia, Departamento de Ecología e Hidrología</i>	
<i>Facultad de Biología, Sec. Apoyo Estadístico. Servicio de Apoyo a la Investigación (SAI)</i>	
<b>Autores e Instituciones</b>	<b>41</b>
<b>Índice de autores</b>	<b>42</b>
<b>Índice de Instituciones</b>	<b>43</b>



# **Información General**

# Presentación

---

Las V Jornadas de Usuarios de R tendrán lugar en el [Etopia-Centro de Arte y Tecnología de Zaragoza](#), los días 12 y 13 de diciembre de 2013. Etopia es un centro de creatividad, innovación y emprendimiento, 16.000 m<sup>2</sup> para el trabajo colaborativo, la búsqueda de nuevos caminos y para aprender haciendo y compartiendo. Forma parte de la [Milla Digital de Zaragoza](#), promovida por el [Área de Tecnología del ayuntamiento de Zaragoza](#).

Las jornadas, como no podría ser de otra forma, van a incluir trabajos de todos los ámbitos y están abiertas tanto a usuarios como a entusiastas de R independientemente de su área de interés. Los objetivos para estas jornadas serán los mismos que para las anteriores que tan buenos resultados obtuvieron. Estos objetivos incluyen:

- Proporcionar un punto de encuentro a los usuarios de R
- Fomentar la colaboración entre ellos en un ambiente multidisciplinar
- Divulgar el conocimiento del lenguaje y sus posibilidades
- Promover el uso de R

Usuarios y entusiastas de R de todos los ámbitos —universidad, institutos de investigación, administraciones públicas, empresa privada— están invitados a participar en las V Jornadas y compartir con la comunidad aplicaciones y ejemplos interesantes que reflejen la madurez de R y la diversidad de los problemas y campos en los que viene utilizándose con éxito. Existen las siguientes modalidades de participación:

- Comunicaciones orales de 15 minutos seguidas de una discusión de 5 minutos (la decisión sobre la duración podría sufrir modificaciones en función del número final de ellas).
- Presentaciones breves de 5 minutos, donde el ponente expone en tres diapositivas (número orientativo) de forma breve y concisa, quién es/son, qué ha/n hecho, y qué resultados y conclusiones se extraen de ello que puedan ser de interés para otras personas.
- Talleres de 2 horas aproximadamente, donde se explican paquetes, procedimientos, y programas de R. En esta edición, además de las ponencias invitadas, las presentaciones orales y los talleres, se llevarán a cabo presentaciones breves donde el ponente expondrá de forma concisa los resultados y conclusiones de alguna investigación llevada a cabo con R que puedan ser de interés para otros colegas.

---

Desde el comité organizador nos gustaría destacar la excelente labor llevada a cabo por el comité científico, a los ponentes de los talleres y a todos los asistentes y patrocinadores que han permitido confeccionar el programa que a continuación detallamos y esperamos que sea de vuestro interés.

Esperamos que las jornadas resulten lo más provechosas posibles y que disfrutéis de una confortable estancia en Zaragoza.

# Información útil

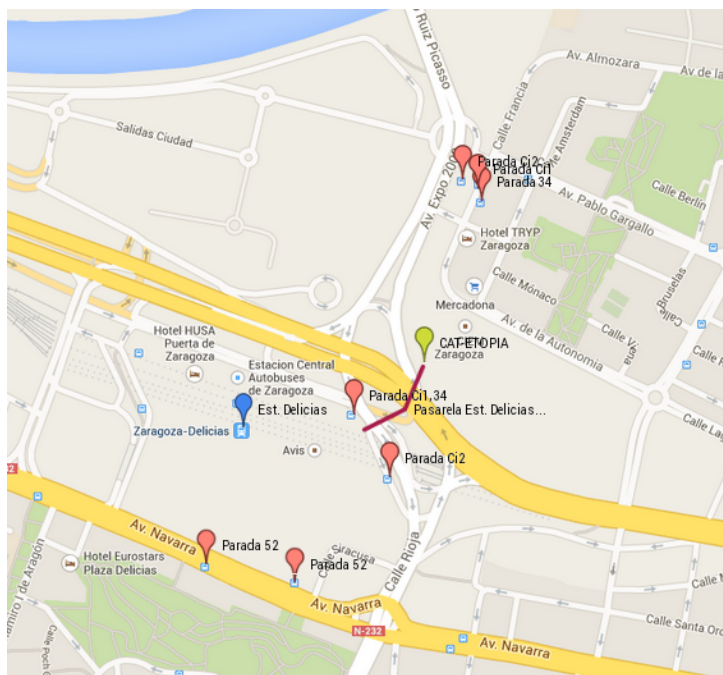
---

## Ubicación de las jornadas

Las jornadas se celebrarán en el Centro de Arte y Tecnología [CAT-ETIOPIA](#) que el Ayuntamiento de Zaragoza ha desarrollado en la llamada [Milla Digital](#).



El C.A.T. está situado prácticamente en el centro de la Milla Digital, justo enfrente de la Estación de Delicias, con la cual se encuentra comunicado mediante un puente peatonal conocido como la [Pasarela Delicias](#). Este área de Zaragoza se encuentra comunicado con el centro de la ciudad (Puerta del Carmen y Plaza del Paraíso) mediante las líneas de autobuses [34](#) y [52](#), también pueden resultar de interés las líneas circulares [Ci1](#) y [Ci2](#).



Las comunicaciones orales y breves se llevarán a cabo en el auditorio William J. Mitchell del CAT-ETIOPIA. Para acceder al edificio cada participante se deberá identificar en recepción donde disponen de una lista con todos los asistentes.

## **Talleres**

Los participantes a los talleres deben traer su propio ordenador portátil con las herramientas y en las condiciones que indiquen los responsables de los talleres. La inscripción a los talleres se realizará en el momento de la entrega de material, en el momento de la recepción de los participantes. Dado el limitado número de plazas, se reservará plaza por orden de inscripción. Los talleres se desarrollarán en las aulas de que dispone el CAT-ETOPÍA para laboratorios audiovisuales y meetings.

## **Certificados**

Los certificados relativos a la asistencia y participación en las jornadas se entregarán durante la celebración de las mismas.

## **Material**

Todo el material estará disponible a través de la página web de las Jornadas .

# Comité organizador

---

- [Carlos Gil Bellosta](#) (coordinador, Datanalytics)
- [Sergio Jiménez](#) (Scien Analytics)
- Luis Mariano Esteban (Universidad de Zaragoza)
- [Rubén Moreno Ruíz](#) (Scien Analytics)
- [Miguel Ángel Luzón](#) (Scien Analytics)
- Jorge Ojeda (Universidad de Zaragoza)
- [Xavier de Pedro Puente](#) (Vall d' Hebron Research Institute)
- Emilio Torres Manzanera (Universidad de Oviedo)

# Comité científico

---

- Sandra Barragán (Universidad de Valladolid)
- Ramón Díaz-Uriarte (Universidad Autónoma de Madrid)
- Juan Ramon González (Centro de Investigación en Epidemiología Ambiental)
- [Oscar Perpiñán](#) (Universidad Politécnica de Madrid)
- Miguel Angel Rodríguez (coordinador, Xunta de Galicia)
- Isaac Subirana (Instituto Hospital del Mar de Investigaciones Médicas)
- Joan Vila (Instituto Hospital del Mar de Investigaciones Médicas)
- [Otto F. Wagner](#) (Ilustre Colegio de Economistas de Madrid)

# Patrocinadores

---





# Programa

---

## ■ JUEVES 12 DE DICIEMBRE

- Sesión de mañana:

9.00-9.30 Recepción y entrega de material

9.30-10.00 Inauguración de las jornadas

10.00-11.00 Conferencia plenaria:

"adabag: An R Package for Classification with Boosting and Bagging" -  
*Esteban Alfaro Cortés, Matías Gámez y Noelia García*

11.00-11.30 Pausa café

11.30-13.30 Sesión de Comunicaciones

- ◇ Comunicaciones orales

- Package xkcd: Plotting ggplot2 graphics in a XKCD style (*Emilio Torres-Manzanera*)

- Métrica de Wasserstein para la comparación de matrices origen-destino (*Aleix Ruiz de Villa*)

- Categorización automática de contenidos web con R (*Pedro Concejero*)

- Algunos aspectos prácticos del manejo de datos de encuesta con R (*Jesús Bouso Freijo*)

- Sesgo de publicación en ciencias médicas (*Borja Santos Zorrozuía*)

- Utilidad clínica de modelos predictivos: análisis mediante funciones de densidad de probabilidad estimadas por métodos tipo kernel (*Luis Mariano Esteban*)

- Evaluación del uso de modelos mixtos para estimación de la tasa de paro con poca muestra (*José Luis Cañadas Reche*)

- ◇ Presentaciones breves

- Mejora de la detección visual de datos atípicos mediante una modificación en las caras de Chernoff (*Beatriz González Pérez*)

- Tratamiento de datos con R para control de calidad basado en valoraciones biológicas. Rectas Paralelas (*Faustino Huertas Muñoz*)

- Análisis exploratorio de datos del mercado eléctrico español con R (*J.M. Velasco*)

13.30-14.00 Ponencia invitada:

"Optimización Entera Mixta No Lineal (MINLP) con R y Pyomo: Un ejemplo práctico" – *Jorge Ayuso Rejas*

- Sesión de tarde:

16.00-17.00 Conferencia plenaria:

"Mejora de la calidad con R: Aplicación de Seis Sigma y otros métodos estadísticos" – *Emilio López Cano*

17.00-19.00 Talleres paralelos

- "Big data analytics: R + Hadoop" – *Carlos Gil Bellosta*

- "Relenium, selenium en R. Un nuevo paquete para webscraping" – *Aleix Ruiz de Villa*

19.00-20.00 Asamblea "Comunidad R-Hispano"

■ **VIERNES 13 DE DICIEMBRE**

- Sesión de mañana:

9.00-10.00 Conferencia plenaria:

"Análisis de datos reproducible con R: métodos, herramientas y tendencias" – *Felipe Ortega*

10.00-11.00 Mesa redonda: *Retos "para y desde" R*

- ◇ Moderador: *Emilio Torres Manzanera*

- ◇ *Jesús Bouso* (Centro de Investigaciones Sociológicas)

- ◇ *Santiago Basaldúa* (Synergic Partners)

- ◇ *Xavier de Blas* (U. Ramon Llul)

- ◇ *Carlos Gil Bellosta* (R-Hispano)

- ◇ Representante Open Data Ayuntamiento de Zaragoza

11.00-11.30 Pausa café

Punto de encuentro profesional. Cristina Guirado, responsable de recursos humanos de Synergic Partners recogerá curriculum vitae a las personas interesadas.

11.30-13.30 Sesión de Comunicaciones

- ◇ Comunicaciones orales

- Desarrollo de Interfaces Web utilizando programación funcional en R (*Jorge Luis Ojeda Cabrera*)

- Previsión de equipamientos educativos, culturales y sanitarios en los barrios de nueva creación de la ciudad de Zaragoza (*Sergio Jiménez Sanjuán*)

- El paquete W2CWM2C: análisis de correlación de wavelet. Casos bivariado y multivariado (*Josué M. Polanco Martínez*)

- Análisis automatizado de cuasi-implicaciones. El Proyecto RCHIC: primeros pasos (*Rubén Pazmiño*)

- Postprocesado de resultados de analysis de elementos finitos con R (*Andres Sanz-Garcia*)

- Medición de la potencia en deportistas usando R y encoders (*Xavier de Blas Foix*)

- Estrategias de captación de clientes en mercados con competencia (*Francisco Jesús Rodríguez Aragón*)

- ◇ Presentaciones breves

- 
- Preprocesado de imágenes hiperespectrales en R (*Rubén Urraca Valle*)
  - Análisis clasificatorio de la actividad electroencefalográfica a través del paso de señales temporales al dominio de la frecuencia (*Roberto Fernandez Martinez*)
  - Simulación en R de modelos definidos en hoja de cálculo (*Ramiro Serrano-García*)

13.30-14.00 Ponencia invitada:

"Aplicaciones de Big Data en R" — *Synergic Partners*

- Sesión de tarde:

16.00-17.00 Conferencia plenaria:

" Inferencia Bayesiana para modelos espaciales y espacio-temporales con R" – *Virgilio Gómez Rubio*

17.00-19.00 Talleres paralelos

- "Cazando información espectro-temporal en datos ambientales con R" – *Josué M. Polanco Martínez*
- "Docencia de R mediante investigación reproducible. RStudio, knitr, markdown" – *José Antonio Palazón Ferrando*

19.00-19:30 Clausura Jornadas



**I**

## **Sesión de Comunicaciones Jueves**

# 1 Evaluación del uso de modelos mixtos para estimación de la tasa de paro con poca muestra

---

*José Luis Cañadas Reche*

*Instituto de Estudios Sociales Avanzados (IESA-CSIC)*

La EPA, a pesar de ser la mayor encuesta de España, no ofrece muestra suficiente para algunas desagregaciones, tal es el caso por ejemplo, si queremos estimar la tasa de paro de los hombres de 35 a 40 años residentes en Zaragoza y con estudios universitarios.

El uso de modelos mixtos se ha utilizado tradicionalmente para modelar estructuras de covarianzas no contempladas por los modelos lineales tradicionales. Los modelos mixtos, sin embargo, también pueden ser utilizados para obtener unas estimaciones más precisas de las medias condicionales.

Para comprobarlo, se utilizó R para comparar la estimación clásica con la obtenida mediante modelos mixtos. Se tomaron diversas 5 submuestras de la EPA de diferente tamaño. Se calculó la tasa de paro a nivel provincial mediante ambos métodos repitiendo el proceso 200 veces, obteniendo como medida de precisión el error absoluto medio. Los modelos mixtos dieron un menor EAM incluso para muestras inferiores al 5 % de la encuesta.

## **Bibliografía**

1. Gelman, A. y Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge New York: Cambridge University Press.
2. Lax, J.R y Philips, J.H (2009). How Should We Estimate Public Opinion in The States?. American Journal of Political Science, Vol 53, No 1, pp 107-121

## 2 Algunos aspectos prácticos del manejo de datos de encuesta con R

---

*Jesús Bouso Freijo*  
*Centro de Investigaciones Sociológicas (CIS)*

La presentación pretende ser un breve compendio de algunas herramientas útiles contenidas en diversos paquetes para el manejo de datos de encuesta. Fundamentalmente, las ideas a exponer proceden de la experiencia adquirida trabajando con R en el Centro de Investigaciones Sociológicas (CIS). Los datos de estudios del CIS cuentan con la particularidad de presentar una estructura variable que hace muy complicada la automatización sistemática del manejo de los mismos. También es relevante para su tratamiento con R la supremacía del programa SPSS en el ámbito de la Sociología, las Ciencias Políticas y otras disciplinas sociales afines. Por su parte, Stata va adquiriendo cierta presencia en estos ámbitos. Ello hace conveniente analizar las posibilidades que ofrece R a la hora de interactuar con datos de otros paquetes. Por otra parte, se presenta brevemente el modo en que la batería de series temporales publicada por el CIS denominada “Indicadores del Barómetro” se halla implementada en R. Por último, se introduce muy someramente el papel jugado hasta ahora por R en el tratamiento estándar de metadatos de encuestas.

En resumen, cabe citar como puntos principales a tratar los siguientes:

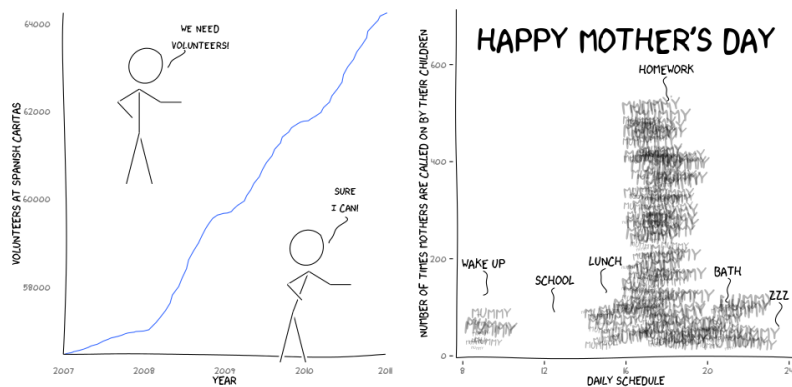
- Interacción con datos de otros paquetes estadísticos
- Interacción con bases de datos
- Ideas para la lectura de ficheros de estructura variable (como los estudios del CIS)
- Utilización de R en el CIS: Los Indicadores del Barómetro
- Metadatos con R: Data Documentation Initiative (DDI)

### 3 Package xkcd: Plotting ggplot2 graphics in a XKCD style

---

*Emilio Torres Manzanera*  
*Universidad de Oviedo*

Se presenta el paquete `xkcd`, que realiza gráficos `ggplot2` como si fueran trazados a mano, siguiendo el estilo de las tiras cómicas de XKCD.





## 4 Métrica de Wasserstein para la comparación de matrices origen-destino

---

*Aleix Ruiz de Villa, Jordi Casas, Martijn Breen*  
*TSS - Transport Simulation Systems*  
*RugBcn, Grupo de usuarios de Barcelona*

Las matrices origen-destino (OD) son un elemento básico en los estudios de tráfico. Dada una red de transporte (por ejemplo una autopista con sus vías secundarias), describen el número de viajes que se dan en un intervalo de tiempo, donde los orígenes y destinos pertenecen a un conjunto fijo de localizaciones, llamados centroides.

El problema que abordamos aquí es el de comparar dos matrices OD. En un principio, se pueden ver las diferencias celda a celda. Sin embargo, esta comparación no recoge la topología del red. Es decir, dos centroides muy cercanos pueden tener viajes muy diferentes, debido por ejemplo a las perturbaciones del proceso de muestreo, pero en esencia ambas matrices recoger el mismo tipo de información.

Para abordar dicho problema, utilizamos técnicas de transporte de masas, una rama teórica de las matemáticas, íntimamente relacionada con problemas de transporte. Dados dos pares  $od$  ( $o_1, d_1$ ) y ( $o_2, d_2$ ), definimos la distancia entre ellos, como el tiempo de transporte (calculado en base a la topología de la red) necesario para desplazarse de un origen al otro y volver del correspondiente destino: es decir  $d(o_1, o_2) + d(d_2, d_1)$ . Bajo estas circunstancias, definimos (informalmente) la distancia entre matrices  $od$ , como el mínimo tiempo de desplazamiento para mover la masa total de la matriz (número total de viajes)  $od_1$  hasta  $od_2$  y luego devolverla. En transporte de masas, esta distancia es conocida como la distancia de Wasserstein. Este problema se resuelve mediante técnicas básicas de programación lineal.

El principal interés de este método, es que creemos que se puede utilizar en otras áreas científicas como el estudio de movimientos demográficos o el estudio de redes de telecomunicaciones y que podría tener aplicaciones peculiares como la comparación de ofertas de vuelo de dos compañías aéreas. Para ello desarrollamos un paquete en R, que permita fácilmente el cálculo de dicha distancia.

### Bibliografía

1. `lp_solve` and Kjell Konis. (2013). `lpSolveAPI`: R Interface for `lp_solve` version 5.5.2.0. R package version 5.5.2.0-8. <http://CRAN.R-project.org/package=lpSolveAPI>
2. Villani, C. (2003) Topics in optimal transportation. American Mathematical Society, Providence.

## 5 Mejora de la detección visual de datos atípicos mediante una modificación en las caras de Chernoff

---

*Beatriz González Pérez, Victoria López López, Jorge Cordero  
Universidad Complutense de Madrid*

En este trabajo se realiza una mejora de la función de R que construye el gráfico de las caras de Chernoff para un perfil multivariante. Esta mejora se realiza mediante una categorización utilizando una paleta de colores y se aplica a una base de datos real. El procedimiento proporciona al investigador una mayor capacidad visual a la hora de detectar datos atípicos.

## 6 Categorización automática de contenidos web con R

---

*Pedro Concejero, César García, Ana Armenta, Paulo Villegas, J. Gregorio Escalada  
Telefónica Digital, Product Development and Innovation*

Telefónica Digital – PDI ha desarrollado un diccionario de contenidos web tomando como base la jerarquía temática y las clasificaciones del Open Directory Project, también conocidas como DMoz –por [directory.mozilla \(http://www.dmoz.org/\)](http://www.dmoz.org/). Se trata de un proyecto colaborativo abierto y multilingüe, en el que editores voluntarios listan y categorizan enlaces a páginas web. Muchos creadores de contenidos web categorizan los mismos en dmoz con el fin de obtener una buena posición en los buscadores, pues muchos de ellos utilizan este directorio como semilla para realizar el crawling de Internet completo

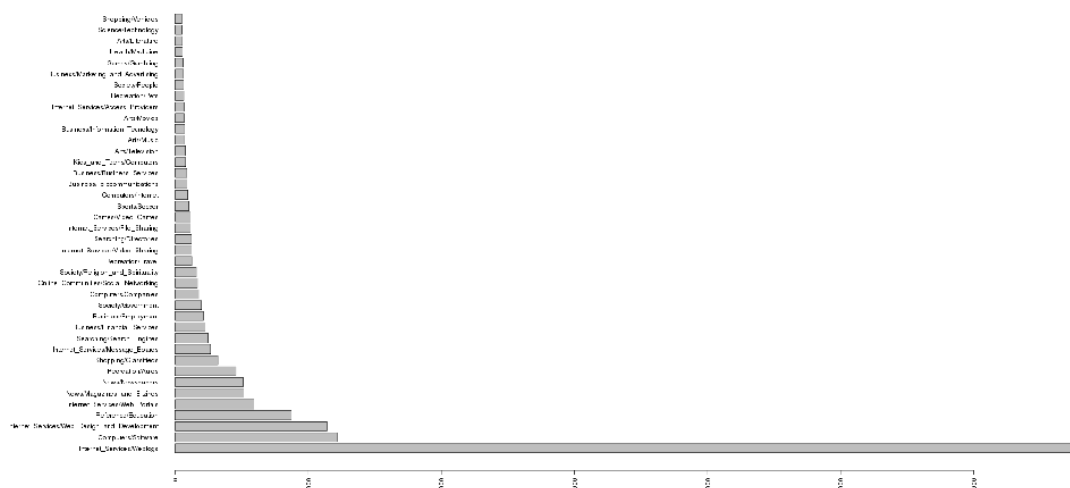
Dos limitaciones importantes de esta taxonomía son su cobertura limitada, esto es, el contenido que no ha sido clasificado en DMoz, y su estructura desbalanceada (la profundidad de la jerarquía y su densidad es muy variable por categorías). Resulta por tanto interesante plantearse un proceso que pueda proporcionar la categoría o clasificación de un contenido web de forma automática, tomando como input el texto completo obtenido de webs reales mediante un crawler, sobre un subconjunto más balanceado de la jerarquía del ODP. Esta presentación describirá el proceso completo que comienza con el análisis de logs representativos de navegación web de usuarios, con el objetivo de seleccionar las categorías más populares o significativas, para luego extraer automáticamente el contenido (texto) completo de las páginas webs asignadas a estas categorías.

La extracción de contenido web (crawl) se realizó mediante nutch (un módulo de apache), al que se le pasaron un total de 84000 dominios que tienen al menos un visitante en un periodo de tiempo. Sin embargo, no podemos extraer automáticamente el texto de todos los dominios que le pasemos, debido a errores tipo “Forbidden” (la web destino no permite la extracción de texto) o “Service unavailable” (el servidor web destino no funciona). El proceso de extracción de texto, configurado con profundidad 1 (sólo se extrae la página principal) obtiene 62748 documentos.

Un proceso de identificación de idioma –mediante tecnología desarrollada por el grupo de Tecnología del Habla de Telefónica I+D- que permite identificar al menos castellano, catalán, euskera, gallego, inglés, francés y portugués, además de proporcionar una medida de confianza de la identificación. Para esta primera fase del proyecto hemos elegido aquellos contenidos en castellano y con una medida de confianza mayor que 0. Este proceso produce un conjunto de 16700 documentos que serán el input para el pre-proceso de texto con la librería R tm. Después de una limpieza básica de caracteres de puntuación y otros especiales, se eliminan las “stopwords” (preposiciones, conjunciones, palabras que no añaden significado) y en esta fase del proyecto se

ha utilizado el “stemmer” (o lematizador, que permite extraer la raíz de una palabra para mantener un único ejemplar para masculino y femenino, o plural/singular) que implementa tm para R3.0 mediante la librería SnowBallC.

La figura a continuación muestra el número de dominios finalmente disponible para entrenamiento por categoría DMOZ. Se observa el enorme desequilibrio entre la categoría más frecuente, que es la de “Internet\_Services-Weblogs” (o páginas personales) que suponen 10000 de los 16000 dominios que son input al clasificador. Ningún clasificador funciona bien con este gran desequilibrio, por lo que finalmente se hizo un muestreo aleatorio de esta categoría del 25 %, fijando el conjunto de dominios input al procedimiento de entrenamiento y validación del clasificador en 10000 documentos en 49 categorías, con un número mínimo de 40 dominios por categoría. Además se filtran palabras poco frecuentes, dejando la matriz de input en 10000 documentos x 3874 palabras.



Los conjuntos creados se aplican a los algoritmos de clasificación incluidos en la librería RtextTools. RtextTools facilita todo el proceso al crear un objeto R (denominado container) que contiene los conjuntos de entrenamiento y validación, en proporciones 70/30, respectivamente. RtextTools también facilita enormemente el proceso de entrenamiento, puesto que incluye funciones por defecto para entrenar hasta nueve clasificadores, y la medición posterior de precisión y otros indicadores de rendimiento de cada uno de los algoritmos probados.

Los clasificadores disponibles en RtextTools v. 1.4.1. (de fecha agosto 2013) y que proporcionaron resultados (funcionaron) son: SVM (support vector machine, Meyer et al., 2012), máxima entropía (o regresión logit multinomial, Jurka, 2012), SLDA (scaled linear discriminant analysis, Peters and Hothorn, 2012), random forests (Liaw and Wiener, 2002). Los siguientes algoritmos no funcionaron, por distintos motivos: GLMNET (Friedman et al, 2010), redes neuronales (Venables y Ripley, 2002), BAGGING (Peters and Hothorn, 2012) y BOOSTING (Tuszyński, 2012). El algoritmo de árboles de decisión dio error porque no admite más de 32 categorías objetivo (Ripley, 2012).

El proceso se realizó en un servidor RedHat 6 con R3.0, con procesador Quad-Core AMD Opteron y 16 GB de RAM. La tabla que presentamos a continuación muestra la proporción promedio (para todas las categorías) con la que los dos mejores algoritmos predicen que un dominio del conjunto de validación pertenece a la clase en la que

---

realmente está clasificado. Esto es, proporción de clasificaciones correctas promediada para todo el conjunto de textos contenido en el conjunto de validación.

- Support Vector Machines (SVM) – 13.4 minutos – 0.62 (proporción de aciertos promedio) - Maximum Entropy (MAXENT) – casi 2 horas – 0.61 (proporción de aciertos promedio)

Estos resultados mejoran sensiblemente si eliminamos las categorías más minoritarias, que fueron reagrupadas en categorías genéricas. Por ejemplo, todos los deportes que no fueran fútbol (“Sports-Soccer”) fueron introducidos al clasificador como “Sports-Others”). Si ignoramos estas categorías, la precisión promedio de SVM sube a 0.72 en ambos casos.

La principal conclusión es que el entrenamiento de un algoritmo como SVM, incluso con sus opciones por defecto, tal y como lo proporciona RtextTools es suficientemente prometedora tanto en eficiencia como en precisión para continuar el desarrollo y optimización del clasificador para un conjunto muy amplio de categorías. Y también que la librería RtextTools facilita enormemente todo el proceso para elaborar un prototipo de este complejo sistema y tomar decisiones sobre los siguientes pasos.

## Bibliografía

Feinerer, I. (2010). Introduction to the tm Package Text Mining in R, 1-7. Retrieved from <http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf>

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010 URL <http://www.jstatsoft.org/v33/i01/paper> Ingersoll, G. S., Morton, T. S., & Farris, A. L. (2013). *Taming Text: How to find, organize and manipulate it*. New York: Manning.

Jurka, T. P., (2012). maxent: An R package for low-memory multinomial logistic regression with support for semi-automated text classification. *The R Journal*, 4(1):56-59, June 2012. URL [http://journal.r-project.org/archive/2012-1/RJournal\\_2012-1\\_Jurka.pdf](http://journal.r-project.org/archive/2012-1/RJournal_2012-1_Jurka.pdf)

Jurka, T. P., Collingwood, L., Boydston, A. E., Gross-, E., & Atteveldt, W. Van. (2013). Package RTextTools .Retrieved from <http://cran.r-project.org/web/packages/RTextTools/RTextTools.pdf>

Jurka, T. P., Collingwood, L., Boydston, A. E., Grossman, E., & Van, W. (2013). RTextTools : A Supervised Learning Package for Text Classification. *The R Journal*, 5, 6-12. Retrieved from <http://journal.r-project.org/archive/2013-1/collingwood-jurka-boydstun-et-al.pdf>

Liaw, A. and M. Wiener (2002). Classification and regression by randomForest. *R News*, 2(3):18-22, 2002. URL [http://www.r-project.org/doc/Rnews/Rnews\\_2002-3.pdf](http://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf)

Meyer,D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch (2012): e1071 Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-1. URL <http://CRAN.R-project.org/package=e1071>

Peters, A. and T. Hothorn (2012). ipred: Improved Predictors, 2012. URL <http://CRAN.R-project.org/package=ipred> . R package version 0.8-13

Qi, X., & Davison, B. D. (2009). Web Page Classification : Features and Algorithms \*. *ACM Computing Surveys*, 41(June), 1-31. Retrieved from <http://www.cse.lehigh.edu/~xiq204/pubs/classification-survey/LU-CSE-07-010.pdf>

Radovanovic, M., & Ivanovi, M. (2008). TEXT MINING : Bag-of-Words Document Representation Machine Learning with Textual Data. Novi Sad Journal of Mathematics, 38(3), 227-234.

Ripley, B. (2012). tree: Classification and Regression Trees, 2012. R package version 1.0-31. URL <http://CRAN.R-project.org/package=tree>

Tuszynski, J. (2012). caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc., 2012. R package version 1.13. URL <http://CRAN.R-project.org/package=caTools>

Venables, W. and B. Ripley (2002): Modern Applied Statistics with S. Springer, New York, fourth edition, 2002

## 7 Sesgo de publicación en ciencias médicas

---

*Borja Santos Zorrozuía, Eduardo González Fraile, Javier Ballesteros Rodríguez*  
*Universidad del País Vasco (UPV/EHU)*  
*Cibersam*  
*Programa PREDOC Gobierno Vasco*  
*Instituto de Investigaciones Psiquiátricas*

El metaanálisis es un herramienta muy utilizada en las ciencias médicas para relai-  
zar una síntesis de la evidencia científica publicada relacionada con un mismo tema. A  
pesar de ser una técnica depurada, cuenta con posibles limitaciones y errores sitemáti-  
cos.

El sesgo de publicación supone una de sus mayores limitaciones. Se define como la  
no publicación de manera deliberada de estudios no favorables a las hipótesis estable-  
cidas previamente. Los motivos de este fenómeno pueden ser entre otros: intereses co-  
merciales de medicamentos, falta de interés de publicación por parte del investigador  
independiente, limitaciones idiomáticas o de localización, o limitaciones editoriales.

La existencia de este sesgo se traduce en una estimación errónea del tamaño del  
efecto combinado de varios estudios (los trazados y publicados). Es por esto que exis-  
ten diferentes técnicas para ajustar el tamaño del efecto combinado asumiendo la exis-  
tencia de dicho sesgo.

El objetivo de esta presentación es probar el funcionamiento de las diferentes li-  
brerías existentes en R que permiten ajustar por la existencia de sesgo de publicación:  
meta, metafor, Copas, SAMURAI, selectMeta. Para ello utilizaremos una serie de estu-  
dios que analizan la efectividad de la agomelatina como tratamiento de la depresión.  
Este conjunto está formado por estudios ya publicados (corroboran la eficacia de este  
tratamiento) y de otros que no han sido publicados (debido a sus pobres resultados).

De esta manera como hemos tenido la posibilidad de metaanalizar la totalidad de  
estudios, conocemos el verdadero tamaño del efecto de la agomelatina. Por lo tanto en-  
frentaremos a este, los estimadores del tamaño del efecto obtenidos al poner en práctica  
las librerías mencionadas anteriormente y de este modo conocer cual es su precisión a  
la hora de calcular el tamaño del efecto.

### Bibliografía

1. Metafor: <http://www.jstatsoft.org/v36/i03/>.
2. Copas: [http://journal.r-project.org/archive/2009-2/RJournal\\_2009-2\\_Carpenter et al.pdf](http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Carpenter%20et%20al.pdf)
3. selectMeta: <http://arxiv.org/pdf/1102.4434v2.pdf>
4. SAMURAI: <http://cran.r-project.org/web/packages/SAMURAI/SAMURAI.pdf>

5. Meta: <http://cran.r-project.org/web/packages/meta/meta.pdf>



## 8 Tratamiento de datos con R para control de calidad basado en valoraciones biológicas. Rectas Paralelas.

---

*Faustino Huertas Muñoz, María Victoria Collazo López, Gloria Frutos Cabanillas  
Agencia Española de Medicamentos y Productos Sanitarios (AEMPS)  
Dpto. de Estadística e Investigación Operativa. Facultad de Farmacia. UCM*

En el control rutinario de la actividad de sustancias de origen biológico en preparaciones farmacéuticas, como las enzimas, factores de coagulación, receptores celulares, antibióticos, etc. se utilizan métodos analíticos cuya interpretación puede basarse en modelos matemáticos como el de rectas paralelas, donde comparando las respuestas de un conjunto de preparaciones referencia de actividad conocida ( $P_s$ ) con las de otro conjunto de preparaciones problema cuya actividad se pretende conocer ( $P_t$ ), es posible obtener  $\log(P_s/P_t) = (\beta_s - \beta_t)/\beta$ , que representa a la relación entre las actividad de la referencia ( $P_s$ ) y de la muestra problema ( $P_t$ ) es igual a la diferencia entre las ordenadas en el origen  $\beta_s$  y  $\beta_t$  de las respectivas rectas, dividido por la pendiente de las rectas paralelas.

En el laboratorio de Hemoderivados de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS) el estudio de rectas paralelas se realiza de acuerdo al programa informático Combistats<sup>®</sup>, elaborado y distribuido por el European Directorate for the Quality of Medicines (EDQM). El programa incluye el análisis del modelo de líneas paralelas, el modelo de razón de pendiente, etc.

Como alternativa de cálculo es posible utilizar R con las funciones básicas como la de modelos lineales (`lm`) y análisis de varianza (`aov`), para conocer si los resultados medidos cumplen las condiciones exigibles de linealidad y paralelismo y, de esta forma, aplicar el modelo para calcular el resultado de la preparación problema ( $P_t$ ) con los límites de confianza correspondientes.

Fundamentalmente, la función `lm()` de R permite conocer si el conjunto formado por dos rectas, obtenidas cada una de ensayos independientes, se puede tratar con uno de los 3 casos o modelos posibles: Que sean parte de la misma recta, que sean dos rectas coincidentes en un punto o que sean rectas paralelas. El estudio mediante `aov()` confirma las conclusiones anteriores de `lm` y permite desglosar la varianza en un mayor número de elementos, y, como ocurre con el programa Combistats<sup>®</sup>, conocer si hay alguna limitación que invalide la aplicación del modelo.

En la exposición se muestra el procedimiento aplicado con R en el ANOVA del modelo y la comparación entre los resultados obtenidos mediante el uso de Combistats<sup>®</sup>.

El cálculo de la actividad de la muestra problema y la estimación del intervalo de confianza podrían realizarse con el paquete `mratios` de G. Dilba et al. Para la implementación del uso de R en control de calidad rutinario, en el tratamiento de resultados de valoraciones biológicas, sería necesario incluir un informe de resultados con trazabilidad a los registros de laboratorio.

## Bibliografía

- 1- European Pharmacopoeia 7<sup>a</sup> Ed. Capítulo 5.3. Análisis estadístico de los resultados de las valoraciones y ensayos biológicos.
- 2- Combistats. <http://combistats.edqm.eu/>
- 3- Package `mratios` Ver 1.3.16 Gemechis Dilba Djira, Mario Hasler, Daniel Gerhard, Frank Schaarschmidt en <http://cran.r-project.org/web/packages/mratios/index.html>
- 4- Rstudio <http://www.rstudio.com/>

## 9 Análisis exploratorio de datos del mercado eléctrico español con R

---

*J.M. Velasco, B. González, G. Miñana, R. Caro, H. Marrao, J. Gil, V. López  
Departamento de Arquitectura de computadores y automática. Universidad Com-  
plutense de Madrid.  
Indizen Technologies, S.L.*

En este trabajo se presenta un análisis exploratorio de datos desarrollado con R, aplicado al mercado eléctrico español. Se han utilizado los datos públicos de los años 2011 y 2012 disponibles en [www.omelholding.es](http://www.omelholding.es). En primer lugar se introducen los conceptos necesarios para comprender el mercado de la energía en España, así como las características esenciales sobre este recurso no almacenable. Una vez definidas las variables de interés, se analizan formas para medir tanto la oferta como la demanda y de todo ello se infiere el precio en el mercado. Para poder realizar un modelo matemático correcto, se requiere de un análisis de los datos previo donde se determinen las dependencias entre las variables, las correlaciones, los valores atípicos y la normalidad de las variables. Este estudio se ha realizado con R mediante programas fuentes incluidos en el anexo y funciones específicas de librerías de R que también se enumeran y se comentan en el trabajo. Los resultados son de dos tipos: numéricos y gráficos. La gran cantidad de gráficos ofrecen al lector una mejor visualización de los datos y por tanto una mejor interpretación de los resultados.

# 10 Utilidad clínica de modelos predictivos: análisis mediante funciones de densidad de probabilidad estimadas por métodos tipo kernel

---

*Luis Mariano Esteban, Gerardo Sanz, Ángel Borque, José Rubio Briones  
Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza  
Departamento de Métodos estadísticos, Universidad de Zaragoza  
Departamento de Urología. Hospital Universitario Miguel Servet, Zaragoza  
Departamento de Urología. Instituto Valenciano de Oncología, Valencia*

La validez de un modelo predictivo pasa por el análisis de propiedades tales como su calibración, discriminación y utilidad clínica. La calibración de un modelo puede ser analizada gráficamente, funciones como `val.prob` de la librería `rms` permiten dicho análisis en R. El estudio de la capacidad de discriminación del modelo se obtiene con el análisis de las curvas ROC y el área bajo la curva (AUC) y puede realizarse con librerías como `ROCR` o `pROC`, pero una vez que hemos comprobado que tenemos un buen modelo predictivo, la aplicabilidad real de dichos modelos pasa por un estudio de su utilidad clínica. Una de las materias que ha recibido más atención últimamente en este campo es la creación de grupos de riesgo que faciliten la aplicación de los modelos predictivos en la práctica clínica diaria. La construcción de estos grupos de riesgo se realiza a través de una selección de puntos de corte sobre las probabilidades que proporciona el modelo y está asociada a la aplicación de distintos tratamientos para al paciente en cada caso. Por ejemplo, si tenemos un único punto de corte, los pacientes pueden ser clasificados como de alto o bajo riesgo para probabilidades por encima o debajo de un cierto valor, y una consecuencia práctica es que podrían ser sometidos a cirugía o no dependiendo de si pertenecen al grupo de alto o bajo riesgo. La selección de un punto de corte óptimo está asociada a unos valores deseados de sensibilidad, especificidad, valor predictivo positivo o valor predictivo negativo, todos estos parámetros pueden ser calculados con una librería como `ROCR`, y probablemente una tabla que nos informe de estos parámetros nos sirva para seleccionar un punto de corte. En los últimos años, además han sido definidos otros parámetros como el beneficio neto y las curvas de decisión que nos permiten comparar el beneficio de aplicar distintos modelos predictivos con una misma selección de puntos de corte y son calculables con la función `dca` de R. Aunque todos estos parámetros nos pueden llevar a seleccionar

---

un buen punto de corte, esta selección se realiza en cierta manera a ciegas, perdiéndose el punto de vista clínico del problema. En este punto creemos que es fundamental el estudio de las funciones de densidad de las distintas poblaciones (sana/enferma) a estudio. La estimación de la densidad de probabilidad mediante funciones tipo kernel nos permite un estudio gráfico del problema con R que nos guiará sobre cómo seleccionar el mejor punto de corte y que da una información clínica sobre la utilidad de los modelos predictivos. En este trabajo queremos ilustrar con ejemplos reales aplicados en oncología como el uso de las funciones de densidad estimadas mediante métodos tipo kernel nos guía en la selección de puntos de corte adecuados y nos informa de una manera clara de la utilidad clínica de los modelos predictivos.

## **Bibliografía**

1. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996. 15(4):361-387.
2. Liu X. Classification accuracy and cut point selection. *Stat Med*. 2012;31:2676-86.
3. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*. 2006. 26(6):565 -574.

## **II**

### **Sesión de Comunicaciones Viernes**

# 11 El paquete W2CWM2C: análisis de correlación de wavelet. Casos bivariado y multivariado.

---

*Josué M. Polanco Martínez*

*Instituto de Economía Pública y Dept. de Econometría y Estadística, Universidad del País Vasco*

El objetivo de esta contribución oral es presentar el paquete R W2CWM2C (disponible en CRAN), sus principales características y algunas aplicaciones utilizando algunos índices bursátiles diarios de la zona Euro. Este paquete contiene cuatro funciones que sirven para producir nuevas herramientas gráficas para el análisis de correlación de wavelet (caso bivariado y multivariado) y un conjunto de datos (siete índices bursátiles de la zona Euro). El paquete W2CWM2C está basado en algunas de las funciones gráficas de los paquetes R Waveslim (Whitcher et al., 2000; Whitcher 2012) y Wavemulcor (Fernandez-Macho 2012a; Fernandez-Macho 2012b), pero añade algunas contribuciones gráficas que ayudan a visualizar de mejor manera los resultados obtenidos al aplicar análisis de correlación de wavelet.

## Bibliografía

1. Fernandez-Macho, J. F. Wavelet multiple correlation and cross-correlation: A multiscale analysis of Eurozone stock markets. *Physica A: Statistical Mechanics and its Applications*, 391(4):1097-1104, 2012a.
2. Fernandez-Macho, J. F. wavemulcor: wavelet routine for multiple correlation, 2012b. R package version 1.2. url: <http://cran.r-project.org/web/packages/wavemulcor/index.html>
3. Polanco-Martinez, J. M. The W2CWM2C package is a set of functions to produce new graphical tools for wavelet correlation (bivariate and multivariate cases), 2012. R package version 1. url: <http://cran.r-project.org/web/packages/W2CWM2C/index.html>
4. Polanco-Martinez, J. M. and Fernandez-Macho, J. F. The package W2CWM2C: description, features and applications. Under review (07/2013) in *Computing in Science & Engineering* (Manuscript Number: CiSE-2013-07-0066).
5. Polanco-Martinez, J. M. and Fernandez-Macho, J. F. Empirical analysis of some peripheral EU stock market indices: A Wavelet cross-correlation approach. Under revision (correction) *Physica*

A: Statistical Mechanics and its Applications (Manuscript Number: PHYSA-12867).

6. Whitcher, B. waveslim: Basic wavelet routines for one-, two- and three-dimensional signal processing, 2012. R package version 1.7.1. url: <http://cran.r-project.org/web/packages/waveslim/index.html>

7 . Whitcher, B., Guttorp, P. and Percival, D.B. Wavelet analysis of covariance with application to atmospheric time series. Journal of Geophysical Research-Atmospheres, 105(D11):941-962, 2000.



# 12 Desarrollo de Interfaces Web utilizando programación funcional en R

---

*Jorge Luis Ojeda Cabrera*  
*Dept- Métodos Estadísticos, Univ. de Zaragoza*

Este trabajo muestra el desarrollo de interfaces web para funciones en R mediante las ideas utilizadas en el paquete 'miniGUI'. Tanto en dicho paquete como en este trabajo se propugna el uso de las capacidades de R para desarrollar programación funcional y 'calcular sobre el lenguaje' a fin de disociar el código necesario para desarrollar los cálculos puramente estadísticos del código utilizado en la construcción de la interfaz de usuario. Esto no sólo ayuda al desarrollo rápido de aplicaciones web, sino que permite separar convenientemente y de una forma sencilla la construcción del Interfaz de la funcionalidad estadística, proporcionando además completa flexibilidad a la hora de desarrollar los interfaces.

En este caso se desarrollan Interfaces Web para el usuario (WUI) en HTML para funciones R que permiten la introducción de los datos mediante formularios HTML. El paquete ha sido probado con la utilidad CGI R FastRWeb y con la aplicación web sumo con configuración básica.

El desarrollo de este trabajo se concreta de momento en una versión incompleta del paquete miniHtmlWUI en la que se implementan todas estas ideas junto con algunos ejemplos básicos de la misma.

## **Bibliografía**

1. miniGUI: tkccl quick and simple function GUI. R package, 2012.  
Jorge Luis Ojeda Cabrera. <http://cran.r-project.org/web/packages/miniGUI/>
2. Conception, evolution, and application of functional programming languages.  
Paul Hudak. ACM Comput. Surv., 21(3):359–411, September 1989.

# 13 Análisis Automatizado de Cuasi-Implicaciones el Proyecto RCHIC: primeros pasos

---

*Rubén Pazmiño, Raphael Couturier, Pablo Gregori  
Escuela Superior Politécnica de Chimborazo. Ecuador  
Universida Comte. Francia  
Universidad Jaume I. España*

El chic (por sus siglas en francés Classification HiérarchiqueImplicative et Cohésitive) es el único programa que permite hacer realidad los resultados teóricos del Análisis Estadístico Implicativo. Ésta teoría se ha desarrollado desde los años 70 por el profesor Régis Grasy colaboradores y permite determinar cuasi-implicaciones entre variables y clases de variables. En forma simplificada permite establecer reglas del tipo: Si se observa a, entonces se observa generalmente b. El software chic es un software propietario de origen francés, elaborado por Raphaël Couturier, que trabaja en la plataforma Windows, en 6 idiomas, con una interface sencilla, liviano y que permite los siguientes análisis: árboles de similaridad, grafo implicativo, árbol cohesitivo y reducción. Este trabajo tiene el objetivo de socializar el proyecto Rchic (chic libre basado en R) y sus avances. El proyecto Rchic consiste en diseñar un entorno colaborativo para elaborar una versión libre del software propietario chic basada en el lenguaje estadístico R.

## Bibliografía

1. Regis, Grass. Contribution a l'étude experimentale et a l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathematiques. Rennes : These d'Etat, 1979.
2. Regnier, Jean Claude. 7? Coloquio Internacional de Analisis Estadistico Implicativo (ASI 7). [En linea] 18 de Junio de 2013. [Citado el: 9 de Septiembre de 2013.] <http://sites.univ-lyon2.fr/asi7/>.
3. Gras, Regis y Bailleul, Marc. Primer Coloquio Internacional de Analisis Estadistico Implicativo (ASI 1). [En linea] 23 de Junio de 2000. [Citado el: 9 de Septiembre de 2013.] [http://math.unipa.it/grim/asi/asi\\_00\\_CAEN.htm](http://math.unipa.it/grim/asi/asi_00_CAEN.htm).
4. Spagnolo, Filippo. I COLOQUIO O METODO ESTATISTICO IMPLICATIVO UTILIZADO EM ESTUDOS QUALITATIVOS DE REGRAS DE ASSOCIACAO CONTRIBUICAO A PESQUISA EM EDUCACAO. [En linea] 9 de Julio de 2003. [Citado el: 9 de Septiembre de 2013.] [http://math.unipa.it/grim/asi/asi\\_03\\_brasil.htm](http://math.unipa.it/grim/asi/asi_03_brasil.htm).

---

5. –. Groupe International d'Analyse Statistique Implicative. Actes des Journees - Proceedings - Atti di Convegni. [En linea] 6 de Octubre de 2005. [Citado el: 10 de Septiembre de 2013.] [http://math.unipa.it/grim/asi/asi\\_index.htm](http://math.unipa.it/grim/asi/asi_index.htm).

6. Regnier, Jean Claude. 6? Coloquio Internacional de Analisis Estadistico Implicativo (ASI6). [En linea] 7 de Noviembre de 2012. [Citado el: 13 de Septiembre de 2013.] <http://sites.univ-lyon2.fr/asi6/>.

# 14 Postprocesado de resultados de análisis de elementos finitos con R

---

*Andres Sanz Garcia, Julio Fernandez Cenicerros, Rubén Urraca Valle, Roberto Fernandez Martinez*

*Division of Bioscience. University of Helsinki, Finland*

*EDMANS. Universidad de La Rioja, Spain*

*TELEVITIS. Universidad de La Rioja, Spain*

*Department of Mining and Metallurgical Engineering and Materials Science. University of Basque Country, Spain*

Los avances en las técnicas de simulación numérica y el desarrollo de entornos GUI para el tratamiento de los datos de entrada/salida ha permitido la generación de modelos más realistas [1]. A pesar de ello, el proceso de simular requiere de una serie de detallados pasos que consumen mucho tiempo y recursos. R-project es un lenguaje de programación que ha crecido en flexibilidad y en usos. De hecho, la automatización de tareas para encaminadas a generar flujos de datos procesados es un campo con gran potencial. Mediante el uso de distintos objetos y sus métodos englobados en librerías, R permite reducir los tiempos de procesamiento de repetidas simulaciones [2]. El proceso mediante la generación de scripts que engloban multiple tareas asociadas a cada paso. Algunas de ellas son la generación aleatoria los datos de entrada, ejecución de tareas o subrutinas, control de salidas y generación de gráficas, etc. En esta comunicación se describe un caso aplicado a la simulación de modelos de sólidos continuos mediante el uso del software ABAQUS[3] y el lenguaje de programación Python.

## Bibliografía

1. Fernandez-Cenicerros,J.,Sanz-Garcia,A.,Antonanzas-Torres,F.,dePison,F.J.M.: Multilayer-perceptron network ensemble modeling with genetic algorithms for the capacity of bolted lap joint. HAIS 2012, LNCS 7208 pp. 545-556 (2012)
2. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013)
3. ABAQUS v.6.14. Analysis User's Manual

# 15 Preprocesado de imágenes hiperespectrales en R

---

*Rubén Urraca Valle, Borja Millán, Roberto Fernandez Martinez, Andres Sanz Garcia*

*TELEVITIS. Universidad de La Rioja, Spain*

*Department of Mining and Metallurgical Engineering and Materials Science. University of Basque Country, Spain*

*Division of Bioscience. University of Helsinki, Finland*

En la actualidad, el desarrollo de los sensores hiperespectrales está abriendo numerosas líneas de investigación. Estos sensores, a diferencia de las cámaras convencionales, son capaces de recoger información en múltiples frecuencias dando lugar a la generación de espectros [1]. Con los espectros el número de datos disponible se multiplica, dando lugar a la aparición de cubos de datos. Sin embargo, un análisis apropiado de los mismos permite identificar diversas propiedades de los materiales. Esto ha propiciado que las técnicas hiperespectrales se estén extendiendo a numerosos campos, desde la medicina a la agricultura pasando por la biología. En esta comunicación se busca describir el proceso de importación y preprocesado de datos procedente de los sensores hiperespectrales a R dentro del sector agrícola. Para ello se trabajará con dos tipos de sensores: un sensor NIR puntual (microPHAZIR Analyzer), que genera un único espectro (vector de datos) y una cámara hiperespectral que abarca tanto el rango NIR como el visible y genera un espectro por cada uno de los píxeles recogidos (cubo de datos). Los objetos tratados serán bayas de uva y hojas de diferentes variedades de cepa. Tradicionalmente, los datos son extraídos de la cámara y preprocesados en software muy especializados proporcionados por el propio fabricante del sensor o en software comerciales como Matlab. Sin embargo, cuando se quiere pasar a la fase de postprocesado, se realiza una transferencia de datos a software más especializados en análisis y de mayor disponibilidad como R. En este trabajo se pretende importar directamente los datos desde el sensor a R, eliminando así el uso de software comercial. Para ello se analiza una de las librerías disponibles en R para el tratamiento de espectros, hyperSpec. El objetivo es importar los diferentes formatos generados por los sensores (.txt, .spc, .pdo ...) y guardarlos como objetos hyperSpec para así facilitar la tarea de análisis. Una vez importados se procede al postprocesado de datos, siendo un proceso clave sobre todo en las imágenes de la cámara hiperespectral donde se dispone de más de 1 espectro. El proceso de postprocesado incluye los siguientes pasos: segmentación, eliminación de picos, eliminación de píxeles muertos, aplicación de filtros, calibrado. Con este proceso se consiguen medidas robustas para la posterior fase de análisis sin la necesidad de utilizar software adicionales a R [2].

## **Bibliografía**

1. Grahn, H.F., Geladi, P. (2007) Techniques and applications of hyperspectral image analysis. Wiley
2. Wehrens, R. (2011) Chemometrics with R. Springer

# 16 Análisis clasificatorio de la actividad electroencefalográfica a través del paso de señales temporales al dominio de la frecuencia

---

*Roberto Fernandez Martinez, Rubén Lostado Lorza, Rubén Urraca Valle, Andres Sanz Garcia*

*Department of Mining and Metallurgical Engineering and Materials Science. University of Basque Country, Spain*

*Universidad de La Rioja. Spain*

*Universidad de La Rioja. Spain*

*Division of Bioscience. University of Helsinki, Finland*

Esta comunicación presenta la primera parte del trabajo realizado para clasificar los diferentes estados o sentimientos que una persona puede tener al realizar ciertas acciones. Se muestra cómo mediante la utilización de un EGG (encefalograma) multicanal se pueden clasificar las emociones que una persona tiene al visionar varios videos. Se analizan diferentes estados como pueden ser emoción y sorpresa, felicidad y placer, logro y compromiso, confusión y desconcierto, y aburrimiento. A través del uso de un EGG se obtienen valores que captan las pequeñas señales eléctricas que las células del cerebro humano producen al comunicarse entre ellas. Posteriormente se convierten las señales recogidas por los 14 canales del EGG al dominio de la frecuencia, utilizando las conocidas técnicas de análisis de Fourier y además diferentes tipos de filtros a la hora de adecuar la señal. Las señales recogidas son filtradas para eliminar ruidos y posteriormente obtener las siguientes variables significativas que según la literatura definen los cambios de energía: banda alfa (8-13 Hz), banda delta (0-4 Hz), banda beta (14-60 Hz) y banda theta (4-7 Hz). Una vez conocidos las bandas en cada situación se realiza un análisis de la varianza para conocer como de precisa puede ser la futura clasificación de los diferentes estados. Para ellos cuatro test de análisis de varianza son utilizados: ANOVA, Bartlett test, Brown-Forsyth test y Fligner-Killeen test. Se analizan los cuatro test para cubrir los casos de variables paramétricas, semi-paramétricas y no paramétricas. Con este análisis se confirma si la hipótesis nula puede ser rechazada y además se conoce cuanto de diferentes pueden ser las clases estudiadas.

# 17 Medición de la potencia en deportistas usando R y encoders

---

*Xavier de Blas Foix  
Universitat Ramon Llull  
FPCEE Blanquerna  
Grupo SAFE  
Chronojump-Boscosystem*

La medición de la fuerza en los deportistas se ha realizado tradicionalmente a partir de observar la máxima carga que éstos pueden levantar, sin ir ligado ello a velocidad, aceleración o potencia. En los últimos años han aparecido en el mercado algunos codificadores (encoders) que calculan la potencia para cada carga levantada, siendo un parámetro mucho más relevante en la mayoría de los deportes, y permitiendo conocer si se está entrenando correctamente. Estos encoders tienen un coste económico alto y no son software libre.

En la comunicación se presentan tres modelos de encoder que pueden conectarse a una placa de hardware libre: Chronopic y un firmware y software de captura y gestión libres. Las piezas de software analizan los datos que proceden del encoder usando scripts de R. El conjunto se conecta al software Chronojump, un software libre que desde hace varios años se comunica con R para sus cálculos.

## **Bibliografía**

1. De Blas Foix, F. X. (2012). Proyecto Chronojump-Boscosystem. Herramienta informática libre para el estudio cinemático del salto vertical: medición del tiempo, detección del ángulo de flexión sin marcadores y elaboración de tablas de percentiles.
2. Gonzalez-Badillo, J.G., y Sanchez-Medina, L.S. (2010). Movement velocity as a measure of loading intensity in resistance training. *Int J Sports Med*, 31, 347-352.
3. Padulles, J.M. (2011). Valoración de los parámetros mecánicos de la carrera. Desarrollo de un nuevo instrumento de medición. Tesis doct. Barcelona: INEFEC, Universitat de Barcelona.
4. Pena, J. (2013). El entrenamiento de la condición física en el voleibol. Fundación CIDIDA.
5. Tous Fajardo, J. (1999). Nuevas tendencias en fuerza y musculación.



# 18 Estrategias de Captación de Clientes en Mercados con Competencia

---

*Francisco Jesús Rodríguez Aragón*  
*Associate Professional Risk Manager*

En este trabajo se lleva a cabo un análisis del entorno competitivo de una empresa determinada junto con la elaboración de una estrategia de búsqueda y optimización, geo-referenciada, de clientes teniendo en cuenta los siguientes hitos principales en su desarrollo:

- Localización de los competidores y el establecimiento de áreas geográficas de concentración
- Ubicación de nichos de mercado y definición de zonas de concentración de lo que se va a entender como mercado potencial
- Facilitar la toma de decisiones en cuanto a:
  - La realización o no de acciones comerciales
  - Dónde realizar las anteriores acciones comerciales
  - La posibilidad de llevar a cabo campañas de publicidad y/o marketing (y de sus problemas derivados como localización de postes publicitarios, optimización del buzoneo, etc)

El informe que aquí se presenta ofrece un Análisis de Prospección de Mercados con el que se ofrece un ejemplo de la potencialidad que se podría obtener del uso efectivo de bases de datos como SABI si se le suma la potencialidad del lenguaje R junto con análisis estadísticos en materia de riesgo y análisis de la competencia. Este trabajo está formado por un conjunto de 5 análisis interrelacionados cuya idea principal se basa en la interrelación de la competencia con el mercado potencial dado un determinado cliente, así pues, en el primer paso se procede a realizar un análisis general y relativo de tipo financiero del estatus de la industria y del sector competitivo considerado en sí, para posteriormente localizar de un modo segmentado a la competencia; tras estos pasos, en el tercero se define lo que se entiende por mercado potencial y cómo localizar nichos claves de nuevos clientes, de modo que en un siguiente paso lo se analiza es la distribución de dichos clientes, para finalmente en el último análisis, relacionar las concentraciones de clientes con las de empresas competitivas de modo más o menos segmentado en base a la calidad crediticia del mercado de un modo que finalmente se puedan tomar decisiones acertadas de actuación muy enfocadas al área marketing-comercial, pero manteniendo en todo momento el sentido clave del riesgo asociado a estos nuevos clientes que integran los mercados potenciales y que aquí se construyen y se analizan. Finalmente debe indicarse que el análisis que aquí se realiza va enfocado fundamentalmente a sociedades que publican (y en general tienen obligación de ello) información financiera excluyéndose a los autónomos y a aquellas sociedades que no la emiten

## **Bibliografía**

1. Bivand R. S.; Pebesma E. J.; Gomez-Rubio V. (2008) Applied spatial data analysis with R Springer, New York, ISBN 978-0-387-78171-6
2. Kahle D.; Wickham H. (2013) ggmap: Spatial Visualization with ggplot2 The R Journal Vol. 5/1, June ISSN 2073-4859
3. Kahle D.; Wickham H. (2013) Package ggmap  
URL: <http://cran.r-project.org/web/packages/ggmap/ggmap.pdf>

# 19 Previsión de equipamientos educativos, culturales y sanitarios en los barrios de nueva creación de la ciudad de Zaragoza

---

*Sergio Jiménez Sanjuán*  
*SCIEN Analytics*

El objetivo fundamental del estudio es hacer una previsión de necesidades futuras de equipamientos para el horizonte temporal 2013- 2022 en los barrios de nueva creación de la ciudad de Zaragoza.

El primer objetivo es la estimación de la población futura de los barrios de nueva creación de la ciudad de Zaragoza. Los barrios a estudiar presentan diferentes problemáticas a la hora de analizar su dinámica poblacional por lo que requerirán métodos y técnicas diferenciadas.

El otro pilar del proyecto es determinar la población a la que es capaz de dar servicio un equipamiento. Responderemos a esta cuestión desde un punto de vista práctico. Determinaremos la población típica a la que están dando servicio, en la actualidad, los distintos tipos de equipamientos que abarca el estudio siguiendo estos pasos:

- Calcular las áreas de influencia de los distintos equipamientos
- Calcular la población total, y composición, que vive dentro de cada área de influencia
- Estudiar estadísticamente las distribuciones de población de todas las áreas de influencia y calcular unos intervalos de población típicos a los que están dando servicio los equipamientos en la actualidad

Finalmente utilizaremos un criterio de mínimos respecto a las necesidades futuras. Es decir, supondremos necesarios un número de equipamientos tal que teniendo en cuenta la población prevista a la que daría cobertura cada equipamiento se situara entre el percentil 75 y 90 de los que atienden a mayor número población en la actualidad (2012).

El objetivo de la ponencia, además de la presentación de los resultados del estudio, es ilustrar el uso de R y de los diferentes paquetes que se ha realizado en su desarrollo:

- Desarga y análisis de datos INE: paquete pxR
- Procesado de cartografías manzana a manzana: PBSmapping, maptools

- Descarga de datos de equipamientos: RJSON, XML
- Cálculo de áreas de influencia de equipamientos: PBSmapping, rgdal
- Análisis de Datos de población y previsión de población futura
- Previsión de población por franjas de edades
- Mapas: ggmap

## Bibliografía

1. Estadística INd (????). Proyeccion de la Poblacion de Espana a Corto Plazo (2011-2021). Metodologia..
2. Viciano FJ, Gil Bellosta C and Perpignan Lamigueiro O (2011). pxR: PC-Axis with R. R package version 0.24, <URL: <http://CRAN.R-project.org/package=pxR>>.
3. Schnute JT, Boers N and Haigh. R (2012). PBSmapping: Mapping Fisheries Data and Spatial Analysis Tools. R package version 2.62.34, <URL: <http://CRAN.R-project.org/package=PBSmapping>>.
4. Keitt TH, Bivand R, Pebesma E and Rowlingson B (2012). rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.7-12, <URL: <http://CRAN.R-project.org/package=rgdal>>.
5. Kahle D and Wickham H (2012). ggmap: A package for spatial visualization with Google Maps and OpenStreetMap. R package version 2.1, <URL: <http://CRAN.R-project.org/package=ggmap>>.

## 20 Simulación en R de modelos definidos en hoja de cálculo

---

*Ramiro Serrano García, Gregorio R. Serrano  
Keller Graduate School of Management  
Universidad Complutense de Madrid*

Presentamos un complemento de Excel para realizar simulación de Montecarlo en R sobre modelos definidos en hoja de cálculo. Con la aplicación (Stochastic-e) se identifican y gestionan las variables del modelo, se definen los parámetros de la simulación y el conjunto de resultados. En cambio, es en R donde se generan los números aleatorios y se realizan los cálculos y análisis estadísticos definidos por el usuario antes de ser devueltos a la hoja de cálculo. Utilizamos el paquete XLConnect, lo que permite adecuar Stochastic-e para su uso con otras hojas de cálculo. Con esta estrategia, el coste de aprendizaje se reduce y la herramienta es accesible para estudiantes de distintas disciplinas mientras se mantiene un elevado nivel de rigor estadístico.

### Bibliografía

1. Baier T, Neuwirth E and Meo MD (2011). Creating and Deploying an Application with (R)Excel and R."The R Journal, \*3\*(2), pp. 5-11.
2. Davis FD (1989). "Perceived usefulness, perceived ease of use and user acceptance of information technology."MIS Quarterly, \*13\*(3), pp. 319-340. <URL: <http://www.jstor.org/pss/249008>>.
3. Heiberger RM and Neuwirth E (2009). R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics. Springer, New York.
4. McCullough BD and Wilson B (2002). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP.Computational Statistics and Data Analysis, pp. 713-721.
5. Seila AF (2005). "Simulation Conference, 2005 Proceedings of the Winter."In Georgia Univ, pp. 7803-9519.

# **III**

## **Talleres**

## 21 Relenium, selenium en R. Un nuevo paquete para webscraping.

---

*Aleix Ruiz de Villa, Lluís Ramon, Andreu Vall  
TSS - Transport Simulation Systems  
RugBcn, Grupo de usuarios de Barcelona*

Actualmente, los paquetes más utilizados para hacer web scraping con R son XML y RCurl. Ambos permiten 'parsear' el código html de la página web y extraer la información que nos interese. Sin embargo, ninguno de ellos permite interactuar con los elementos javascript de la página. Por tanto aquella información que dependa de la ejecución de comandos javascript (por ejemplo, abrir una ventana con una dirección url desconocida, o seleccionar elementos en un menú desplegable) queda inaccesible.

Relenium es un importador del módulo Selenium de java, via rJava. Selenium nació para el testeo automático de páginas web. La diferencia principal con los paquetes descritos anteriormente es que Relenium puede emular la navegación de un usuario humano, es decir, apretar botones, seleccionar menús, etc. El resultado es una navegación por la web intuitiva y sencilla.

En este taller, introduciremos los elementos básicos del lenguaje html y los xpaths, y mostraremos las funcionalidades básicas del paquete reelenium. Lo complementaremos con las funcionalidades básicas de XML. No es necesario ningún conocimiento previo.

### Bibliografía

1. Relenium (<https://github.com/LluísRamon/reelenium>)
2. RCurl (<http://cran.r-project.org/web/packages/RCurl/index.html>)
3. Xml (<http://cran.r-project.org/web/packages/XML/index.html>)

## 22 Big data analytics: R + Hadoop

---

*Carlos J. Gil Bellosta*  
*Datanalytics*

El taller es una introducción al análisis de datos masivos almacenados en Hadoop con R utilizando, principalmente, el paquete `rmr2`. Este paquete permite distribuir tareas paralelizables en distintos nodos para procesar conjuntos de datos que no pueden analizarse en memoria.

Una de las operaciones más básicas que cubrirá el taller es la de contar ocurrencias. Pero también se prestará atención a operaciones más propias de R, tales como construir modelos y realizar predicciones.

Finalmente, se utilizará *hadoop streaming* para realizar simulaciones masivas en paralelo. Este ejemplo servirá, además, para ilustrar los mecanismos internos del paquete `rmr2` y del funcionamiento de Hadoop.

Los asistentes al taller aprenderán qué es Hadoop, las operaciones básicas del sistema de ficheros, a crear sus propios procesos *mapreduce* y, particularmente, a comprender el funcionamiento del sistema de paralelización de tareas.



## 23 Cazando información espectro-temporal en datos ambientales con R

---

*Josué M. Polanco Martínez*

*Instituto de Economía Pública y Dept. de Econometría y Estadística, Universidad  
del País Vasco*

El análisis espectral de wavelet (AEW) vía la transformada continua de wavelet (TCW) es una herramienta muy poderosa para la búsqueda de eventos periódicos, cuasi-periódicos y eventos cuya frecuencia cambia con el tiempo en series temporales ambientales (climatológicas, meteorológicas, hidrológicas, ecológicas, etc.). El AEW es capaz de analizar series temporales no estacionarias (las ambientales suelen serlo), i.e., series cuyas propiedades estadísticas (primer y segundo momento) cambian con el tiempo, es capaz de analizar a la vez en el dominio del tiempo y de la frecuencia y dispone de pruebas de significación estadística. En este taller se presentarán los principios estadísticos necesarios para una adecuada utilización del AEW, tanto para el caso uni como para el bivariado y se enfocará en la interpretación de los resultados. EL AEW se llevará a cabo mediante la utilización de los paquetes R SOWAS (Maraun 2007) y biwavelet (Gouhier y Grinsted 2013).

- Paquetes:

SOWAS: <http://tocsy.pik-potsdam.de/wavelets/>

Biwavelet: <http://biwavelet.r-forge.r-project.org/>

- Objetivo: El objetivo principal de este taller es que la(o)s asistentes sean capaces de analizar sus propios datos ambientales (nótese que aunque el taller se enfoca a este tipo de datos, también es posible analizar otros tipos de datos, teniendo siempre presente las características de los datos a estudio) utilizando análisis espectral de wavelet vía la transformada continua (caso uni y bivariado) haciendo uso de los paquetes R SOWAS y biwavelet. Se invita a los asistentes del taller a traer sus propias series temporales ambientales.
- Duración: Tiempo total: 2 horas
- Especificaciones de software: paquetes SOWAS y biwavelet. Tener instalado R ver. 2.14 (o superior), el paquete SOWAS (primero instale el paquete Rwave -está en CRAN- desde R y después instale desde fuentes el SOWAS, i.e., desde una terminal de GNU/Linux R CMD INSTALL sowas\_0.93.tar.gz, también necesitará tener instalado el paquete stats) y el paquete R biwavelet -también está en CRAN.

Si el taller es aceptado, las personas interesadas en asistir podrían contactarme previamente para la instalación, de todo modos se anexará un HOW TO para la instalación de los paquetes y de las series temporales que se usarán en el taller.

- Conocimientos previos: Saber vagamente lo que es una transformada de Fourier, conocimiento muy elemental de análisis de series temporales, conocimientos básicos de R en línea de comandos.
- Tabla de contenidos:
  1. Breve introducción de conceptos básicos (función wavelet, tipos de funciones wavelet, transformada continua de wavelet, análisis espectral caso uni y bi variado, Fourier vs. wavelet, sobre escalas, octavas y voces, relación entre escalas y frecuencias).
  2. Presentación de los paquetes SOWAS y biwavelet (funciones utilizadas en este taller, diferencias entre SOWAS y biwavelet).
  3. Estimación e interpretación del espectro wavelet caso uni variado (pruebas de significación estadística y ruido de fondo, poder espectral suavizado vs. crudo. Se presentarán algunos ejemplos de como estimar el espectro wavelet con series temporales ambientales reales, se enfocará en cómo utilizar las funciones que estiman el poder espectral -sobretudo como inicializar los parámetros de entrada- y se analizarán de modo básico los resultados).
  4. Estimación del espectro cruzado, la coherencia normalizada de wavelet y el desfase (caso bivariado) entre dos series temporales ambientales (pruebas de significación estadística SOWAS vs biwavelet, espectro cruzado vs coherencia normalizada, interpretación del desfase. Aplicaciones reales a series ambientales, se explicarán de manera breve como iniciar los principales parámetros de entrada de las funciones que se utilizarán para el análisis bivariado y se analizarán de modo básico los resultados).

## Bibliografía

1. Cazelles, B., M. Chavez, D. Berteaux, F. Menard, J. O. Vik, S. Jenouvrier, and N. C. Stenseth. 2008. Wavelet analysis of ecological time series. *Oecologia* 156:287-304.
2. Grinsted, A., J. C. Moore, and S. Jevrejeva. 2004. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics* 11:561-566.
3. Liu, Y., X. San Liang, and R. H. Weisberg. 2007. Rectification of the Bias in the Wavelet Power Spectrum. *Journal of Atmospheric and Oceanic Technology* 24:2093-2102.
4. Maraun, D, J. Kurths and M. Holschneider. 2007. Nonstationary Gaussian Processes in Wavelet Domain: Synthesis, Estimation and Significance Testing. *Phys. Rev. E* 75, 016707.
5. Maraun D. and J. Kurths. 2004. Cross Wavelet Analysis. Significance Testing and Pitfalls, *Nonlin. Proc. Geoph.* 11(4), 505-514.
6. Polanco-Martinez, J. M. 2011. Aplicación de técnicas estadísticas en el estudio de fenómenos ambientales y ecosistémicos, tesis doctoral, 208 pg. Servicio Editorial de la Universidad del País Vasco (UPV/EHU), ISBN: 978-84-9860-812-0

---

<http://www.ehu.es/argitalpenak/images/stories/tesis/Ciencias/8120PolancoMartinezTH.pdf>

7. Polanco, J., U. Ganzedo, J. Saenz, A. M. Caballero-Alfonso and J. J. Castro-Hernandez. 2011. Wavelet analysis of correlation among Canary Islands octopus captures per unit effort, sea-surface temperatures and the North Atlantic Oscillation, *Fisheries Research*, 107(1-3):177-183

8. Rouyer, T., J. M. Fromentin, F. Menard, B. Cazelles, K. Briand, R. Pianet, B. Planque, and N. C. Stenseth. 2008. Complex interplays among population dynamics, environmental forcing, and exploitation in fisheries. *Proceedings of the National Academy of Sciences* 105:5420-5425.

9. Torrence, C., and G. P. Compo. 1998. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society* 79:61-78.

10. Torrence, C., and P. J. Webster. 1998. The annual cycle of persistence in the El Nino/Southern Oscillation. *Quarterly Journal of the Royal Meteorological Society* 124:1985-2004.

11. Veleda, D., R. Montagne, and M. Araujo. 2012. Cross-Wavelet Bias Corrected by Normalizing Scales. *Journal of Atmospheric and Oceanic Technology* 29:1401-1408.

## 24 Docencia de R mediante investigación reproducible. ‘RStudio’, ‘knitr’, ‘markdown’

---

*Jose Antonio Palazon Ferrando, Antonio Maurandi López  
Universidad de Murcia, Departamento de Ecología e Hidrología  
Facultad de Biología, Sec. Apoyo Estadístico. Servicio de Apoyo a la Investigación  
(SAI)*

La utilización de la metodología de enseñanza basada en problemas puede reforzarse, en el caso del uso de R, con la disponibilidad de herramientas para elaborar documentos de calidad y con vocación reutilizable.

La combinación ‘RStudio’, ‘markdown’ y ‘knitr’ proporciona un entorno de trabajo que puede utilizarse con una formación básica y que rinde resultados de interés tanto conceptual como aplicados; aportando, competencias atractivas para los estudiantes.

La metodología se ha ensayado tanto en formación de grado como en máster con buenos resultados y acogida. Una de las claves es la simplicidad del lenguaje de marcas ‘markdown’; puede que este sea, opcionalmente, la puerta de entrada al uso de LaTeX.

En el taller se realizará una introducción al método de trabajo en el aula, utilizando ejemplos de problemas propuestos en clase; se hará especial hincapié en los aspectos formales y las dificultades que presentan los estudiantes al iniciarse con estas herramientas.

Así, mediante ejemplos, veremos las posibilidades que el paquete de Yihui Xie, ‘knitr’, nos ofrece en el campo de la investigación reproducible aplicado a la docencia de R.

## **Autores e Instituciones**

# Índice de autores

---

- Maurandi López, Antonio, 40
- Armenta, Ana, 7
- Ballesteros Rodríguez, Javier, 11
- Borque, Ángel, 16
- Bouso Freijo, Jesús, 3
- Breen, Martijn, 5
- Cañadas Reche, José Luis, 2
- Caro, R., 15
- Casas, Jordi, 5
- Collazo López, María Victoria, 13
- Concejero, Pedro, 7
- Cordero, Jorge, 6
- Couturier, Raphael, 22
- de Blas Foix, Xavier, 28
- Escalada, J. Gregorio, 7
- Fernandez Cenicerros, Julio, 24
- Fernandez Martinez, Roberto, 24, 25, 27
- Frutos Cabanillas, Gloria, 13
- García, César, 7
- Gil Bellosta, Carlos J., 36
- Gil, J., 15
- González Fraile, Eduardo, 11
- González Pérez, Beatriz, 6
- González, B., 15
- Gregori, Pablo, 22
- Huertas Muñoz, Faustino, 13
- Jiménez Sanjuán, Sergio, 31
- López López, Victoria, 6
- López, V., 15
- Lostado Lorza, Rubén, 27
- Mariano Esteban, Luis, 16
- Marrao, H., 15
- Miñana, G., 15
- Millán, Borja, 25
- Ojeda Cabrera, Jorge Luis, 21
- Palazon Ferrando, Jose Antonio, 40
- Pazmiño, Rubén, 22
- Polanco Martínez, Josué M., 19, 37
- Ramon, Lluís, 35
- Rodríguez Aragón, Francisco Jesús, 29
- Rubio Briones, José, 16
- Ruiz de Villa, Aleix, 5, 35
- Santos Zorrozuía, Borja, 11
- Sanz Garcia, Andres, 24, 25, 27
- Sanz, Gerardo, 16
- Serrano García, Ramiro, 33
- Serrano, Gregorio R., 33
- Torres Manzanera, Emilio, 4
- Urraca Valle, Rubén, 24, 25, 27
- Vall, Andreu, 35
- Velasco, J.M., 15
- Villegas, Paulo, 7

# Índice de Instituciones

---

- Agencia Española de Medicamentos y Productos Sanitarios (AEMPS), 13  
Associate Professional Risk Manager, 29
- Centro de Investigaciones Sociológicas (CIS), 3  
Chronojump-Boscosystem, 28  
Cibersam, 11
- Datanalytics, 36  
Departamento de Arquitectura de computadores y automática. Universidad Complutense de Madrid., 15  
Departamento de Métodos estadísticos, Universidad de Zaragoza, 16  
Departamento de Urología. Hospital Universitario Miguel Servet, Zaragoza, 16  
Departamento de Urología. Instituto Valenciano de Oncología, Valencia, 16  
Department of Mining and Metallurgical Engineering and Materials Science. University of Basque Country, Spain, 24, 25, 27  
Dept- Métodos Estadísticos, Univ. de Zaragoza, 21  
Division of Bioscience. University of Helsinki, Finland, 24, 25, 27  
Dpto. de Estadística e Investigación Operativa. Facultad de Farmacia. UCM", 13  
EDMANS. Universidad de La Rioja, Spain, 24  
Escuela Superior Politécnica de Chimborazo. Ecuador, 22
- Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, 16
- Facultad de Biología, Sec. Apoyo Estadístico. Servicio de Apoyo a la Investigación (SAI), 40  
FPCEE Blanquerna, 28
- Grupo SAFE, 28
- Indizen Technologies, S.L., 15  
Instituto de Economía Pública y Dept. de Econometría y Estadística, Universidad del País Vasco, 19, 37  
Instituto de Estudios Sociales Avanzados (IESA-CSIC), 2  
Instituto de Investigaciones Psiquiátricas, 11
- Keller Graduate School of Management, 33
- Programa PREDOC Gobierno Vasco, 11
- RugBcn, Grupo de usuarios de Barcelona, 5, 35
- SCIEN Analytics, 31
- Telefónica Digital, Product Development and Innovation, 7  
TELEVITIS. Universidad de La Rioja, Spain, 24, 25  
TSS - Transport Simulation Systems, 5, 35
- Universida Comte. Francia, 22  
Universidad Complutense de Madrid, 6, 33

Universidad de La Rioja. Spain, 27  
Universidad de Murcia, Departamento de  
Ecología e Hidrología, 40  
Universidad de Oviedo, 4  
Universidad del País Vasco (UPV/EHU),  
11  
Universidad Jaume I. España, 22  
Universitat Ramon Llull, 28