
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




© 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
Información General	VII
Presentación	IX
Información útil	XI
Comité organizador	XII
Comité científico	XIII
Patrocinadores	XIV
Programa	XV
I Comunicaciones Jueves	1
1 Evaluación del uso de modelos mixtos para estimación de la tasa de paro con poca muestra	2
<i>José Luis Cañadas Reche</i> <i>Técnico de Investigación en el Instituto de Estudios Sociales Avanzados (IESA-CSIC)</i>	
2 Package xkcd: Plotting ggplot2 graphics in a XKCD style	3
<i>Emilio Torres-Manzanera</i> <i>Universidad de Oviedo</i>	
3 Desarrollo de Interfaces Web utilizando programación funcional en R	4
<i>Jorge Luis Ojeda Cabrera</i> <i>Dept- Métodos Estadísticos, Univ. de Zaragoza</i>	
4 Métrica de Wasserstein para la comparación de matrices origen-destino	5
<i>Aleix Ruiz de Villa, Jordi Casas, Martijn Breen</i> <i>TSS - Transport Simulation Systems</i> <i>RugBcn - Grupo de usuarios de R de Barcelona</i>	

7	Categorización automática de contenidos web con R	9
	<i>Pedro Concejero, César García, Ana Armenta, Paulo Villegas, J. Gregorio Escalada, Telefónica Digital, Product Development and Innovation</i>	
6	Mejora de la detección visual de datos atípicos mediante una modificación en las caras de Chernoff	8
	<i>Beatriz González Pérez, Victoria López López, Jorge Cordero</i> <i>Universidad Complutense de Madrid</i>	
7	Categorización automática de contenidos web con R	9
	<i>Pedro Concejero, César García, Ana Armenta, Paulo Villegas, J. Gregorio Escalada, Telefónica Digital, Product Development and Innovation</i>	
8	Previsión de equipamientos educativos, culturales y sanitarios en los barrios de nueva creación de la ciudad de Zaragoza	11
	<i>Sergio Jiménez Sanjuán</i> <i>SCIEN Analytics</i>	
II	Comunicaciones Viernes	12
9	Algunos aspectos prácticos del manejo de datos de encuesta con R	13
	<i>Jesús Bouso Freijo</i> <i>Centro de Investigaciones Sociológicas (CIS)</i>	
10	El paquete W2CWM2C: análisis de correlación de wavelet. Casos bivariado y multivariado.	14
	<i>Dr. Josué M. Polanco Martínez</i> <i>(Investigador invitado) Instituto de Economía Pública y Dept. de Econometría y Estadística, Universidad del País Vasco</i>	
11	ANÁLISIS AUTOMATIZADO DE CUASI-IMPLICACIONES EL PROYECTO RCHIC: PRIMEROS PASOS	15
	<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>	
12	Postprocesado de resultados de analysis de elementos finitos con R	16
	<i>Andres Sanz-Garcia, Julio Fernandez-Ceniceros, Ruben 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>	
13	Preprocesado de imágenes hiperespectrales en R	17
	<i>Rubén Urraca Valle, Borja Millán, Roberto Fernandez-Martinez, Andrés Sanz García</i>	

TELEVITIS. 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
 Division of Bioscience. University of Helsinki, Finland

14 Análisis clasificatorio de la actividad electroencefalográfica a través del paso de señales temporales al dominio de la frecuencia 18

Roberto Fernandez Martinez, Ruben Lostado Lorza, Ruben 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

15 Medición de la potencia en deportistas usando R y encoders 19

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

16 Estrategias de Captación de Clientes en Mercados con Competencia 20

Francisco Jesús Rodríguez Aragón
 Doctor en Estadística por la Universidad de Córdoba
 Associate Professional Risk Manager

17 Sesgo de publicación en ciencias médicas 21

Borja Santos Zorrozuá^{1, 2, 3}, Eduardo González Fraile⁴, , Javier Ballesteros Rodríguez^{2, 4}
 1 Universidad del País Vasco (UPV/EHU), 2 Cibersam (G16), 3 Programa PREDOC Gobierno Vasco, 4 Instituto de Investigaciones Psiquiátricas

18 Docencia de R mediante investigación reproducible. 'RStudio', 'knitr', 'markdown' 22

Jose Antonio Palazon Ferrando y Antonio Maurandi López
 Universidad de Murcia. Comunidad R-Hispano.
 Departamento de Ecología e Hidrología. Facultad de Biología.
 Sec. Apoyo Estadístico. Servicio de Apoyo a la Investigación (SAI)

III Talleres 23

19 Relenium, selenium en R. Un nuevo paquete para webscraping. 24

Aleix Ruiz de Villa, Lluís Ramon, Andreu Vall
 TSS - Transport Simulation Systems (<http://www.aimsun.com/wp/>)
 RugBcn - grupo de usuarios de Barcelona (<http://rugbcn.wordpress.com/>)

20 Cazando información espectro-temporal en datos ambientales con R	25
<i>Dr. Josué M. Polanco Martínez</i>	
<i>Investigador invitado, Instituto de Economía Pública y Dept. de Econometría y Estadística, Universidad del País Vasco.</i>	
Autores e Instituciones	27
Índice de autores	28

Información General

TODO: Vas por aquí

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 m² 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 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 [ETOPIA](#) 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. se encuentra comunicado con centro de la ciudad mediante las líneas de autobuses **TODO: Poner líneas de autobuses y mapa de paradas señalando la estación del tryp, etc...**

Las comunicaciones orales y breves se llevarán a cabo en el **Auditorio del CAT ****
TODO: nombre y situación de la sala.

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 que indiquen los responsables de talleres. **TODO: cómo se desarrollará la inscripción a los talleres** La inscripción de los talleres se realizará tal y como indica la web de las jornadas . Dado el limitado número de plazas, se reservará plaza por orden de inscripción. Los talleres se desarrollarán en **Aulas de los laboratorios del CAT ****
TODO: nombre y situación de las salas, ambas situadas en el espacio que reserva ETOPIA para laboratorios audiovisuales y meetings y que serán convenientemente señal.

Certificados

Los certificados se enviarán por correo electrónico una vez pasadas las Jornadas.

Material

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

Comité organizador

- [J. Gil Bellosta](#) (coordinador)
- [Sergio Jiménez](#) (Scien Analytics)
- Luis Mariano Esteban (U. de Zaragoza)
- [Rubén Moreno Ruíz](#) (Scien Analytics)
- [Miguel Ángel Luzón](#) (Scien Analytics)
- Jorge Ojeda (U. de Zaragoza)
- [Xavier de Pedro Puente](#)
- Emilio Torres Manzanera (U. de Oviedo)

Comité científico

- Sandra Barragán
- Ramón Díaz-Uriarte
- Juan Ramon González
- [Oscar Perpiñán](#)
- Miguel Angel Rodríguez (coordinador)
- Isaac Subirana
- Joan Vila
- [Otto F. Wagner](#)

Patrocinadores



Programa

Sesión de Comunicaciones I

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

José Luis Cañadas Reche

Técnico de Investigación en el 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

2 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.

3 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.

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 R 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 ($o1, d1$) y ($o2, d2$), 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 $d(o1, o2) + d(d2, d1)$. 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) $od1$ hasta $od2$ 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.

5 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 10658 dominios que tienen un número mínimo de visitas. 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). De hecho la selección final, esto es, dominios de los que dispondremos de texto completo (con profundidad 1) y de su categoría DMoz, se reduce a 4072. Un proceso de identificación de idioma –mediante tecnología desarrollada por el grupo de Tecnología del Habla de Telefónica I+D- permite seleccionar cuáles de ellos se utilizarán, en principio castellano, catalán y gallego –filtrando por tanto inglés y euskera entre otros. La figura a continuación muestra el número de dominios finalmente disponible para entrenamiento por categoría DMOZ (imponiendo como requisito al menos 10 dominios por categoría), en total 2283 páginas correspondientes a 44 categorías, que sigue la típica distribución de ley de potencia. Sitios web de noticias (periódicos pero también revistas y publicaciones electrónicas, portales y lugares de desarrollo web) son las categorías con mayor número de dominios clasificados como pertenecientes a ellas en el diccionario DMOZ, seguidos por negocios,

sitios de la Administración Pública y la categoría de automóvil. Este texto será después pre-procesado con la librería *tm*, y se utilizará la implementación del algoritmo de Porter de la librería *SnowBall* como stemmer. El objetivo es obtener una matriz de frecuencias de raíces de palabras, así como posiblemente bigramas, por categoría. Estas webs, o dominios, se dividen en un conjunto de entrenamiento y otro de test, de forma aleatoria, en proporción 80/20 respectivamente, para entrenar y validar clasificadores estadísticos. Los conjuntos creados se aplican a los algoritmos de clasificación incluidos en la librería *RTextTools*. *RTextTools* facilita además enormemente la medición de precisión y otros indicadores de rendimiento de cada uno de los algoritmos probados. La comunicación oral presentará todos los resultados obtenidos en este trabajo. La lista a continuación muestra los resultados preliminares de dos de los algoritmos incluidos en *RTextTools*, sin que el texto haya sido tratado por el stemmer debido a problemas técnicos con la librería *Snowball* –que serán solucionados lo antes posible y sin duda antes de la conferencia. Los algoritmos que no están en la lista es porque no son aplicables o han dado error en esta primera prueba (a menudo debido a desbordamiento de memoria). La lista muestra el tiempo necesario para el cómputo del modelo en un servidor MS-Windows Server 2003 x64, con procesador Quad-Core AMD Opteron y 30 GB de RAM, así como la precisión. Esta medida es, más concretamente, la proporción promedio (para todas las categorías) con la que el algoritmo predice 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) – 48.43 segundos – 0.621 (proporción de aciertos promedio) Maximum Entropy (MAXENT) – 22.15 minutos – 0.714 (proporción de aciertos promedio)

REFERENCIAS (en formato estándar): 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/> Ingersoll, G. S., Morton, T. S., y Farris, A. L. (2013). *Taming Text: How to find, organize and manipulate it*. New York: Manning. Jurka, A. T. P., Collingwood, L., Boydston, A. E., Gross, E., y 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., y 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> Qi, X., y 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., y Ivanovi, M. (2008). TEXT MINING: Bag-of-Words Document Representation Machine Learning with Textual Data. *Novi Sad Journal of Mathematics*, 38(3), 227–234.

6 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.

7 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 10658 dominios que tienen un número mínimo de visitas. 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). De hecho la selección final, esto es, dominios de los que dispondremos de texto completo (con profundidad 1) y de su categoría DMoz, se reduce a 4072. Un proceso de identificación de idioma –mediante tecnología desarrollada por el grupo de Tecnología del Habla de Telefónica I+D- permite seleccionar cuáles de ellos se utilizarán, en principio castellano, catalán y gallego –filtrando por tanto inglés y euskera entre otros. La figura a continuación muestra el número de dominios finalmente disponible para entrenamiento por categoría DMOZ (imponiendo como requisito al menos 10 dominios por categoría), en total 2283 páginas correspondientes a 44 categorías, que sigue la típica distribución de ley de potencia. Sitios web de noticias (periódicos pero también revistas y publicaciones electrónicas, portales y lugares de desarrollo web) son las categorías con mayor número de dominios clasificados como pertenecientes a ellas en el diccionario DMOZ, seguidos por negocios,

sitios de la Administración Pública y la categoría de automóvil. Este texto será después pre-procesado con la librería *tm*, y se utilizará la implementación del algoritmo de Porter de la librería *SnowBall* como stemmer. El objetivo es obtener una matriz de frecuencias de raíces de palabras, así como posiblemente bigramas, por categoría. Estas webs, o dominios, se dividen en un conjunto de entrenamiento y otro de test, de forma aleatoria, en proporción 80/20 respectivamente, para entrenar y validar clasificadores estadísticos. Los conjuntos creados se aplican a los algoritmos de clasificación incluidos en la librería *RTextTools*. *RTextTools* facilita además enormemente la medición de precisión y otros indicadores de rendimiento de cada uno de los algoritmos probados. La comunicación oral presentará todos los resultados obtenidos en este trabajo. La lista a continuación muestra los resultados preliminares de dos de los algoritmos incluidos en *RTextTools*, sin que el texto haya sido tratado por el stemmer debido a problemas técnicos con la librería *Snowball* –que serán solucionados lo antes posible y sin duda antes de la conferencia. Los algoritmos que no están en la lista es porque no son aplicables o han dado error en esta primera prueba (a menudo debido a desbordamiento de memoria). La lista muestra el tiempo necesario para el cómputo del modelo en un servidor MS-Windows Server 2003 x64, con procesador Quad-Core AMD Opteron y 30 GB de RAM, así como la precisión. Esta medida es, más concretamente, la proporción promedio (para todas las categorías) con la que el algoritmo predice 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) – 48.43 segundos – 0.621 (proporción de aciertos promedio) Maximum Entropy (MAXENT) – 22.15 minutos – 0.714 (proporción de aciertos promedio)

REFERENCIAS (en formato estándar): 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> Ingersoll, G. S., Morton, T. S., y Farris, A. L. (2013). *Taming Text: How to find, organize and manipulate it*. New York: Manning. Jurka, A. T. P., Collingwood, L., Boydston, A. E., Gross, E., y 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., y 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> Qi, X., y 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., y Ivanovi, M. (2008). TEXT MINING: Bag-of-Words Document Representation Machine Learning with Textual Data. *Novi Sad Journal of Mathematics*, 38(3), 227–234.

8 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

Sesión de Comunicaciones II

9 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)

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

*Dr. Josué M. Polanco Martínez
(Investigador invitado) 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.

11 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.

12 Postprocesado de resultados de analysis de elementos finitos con R

Andres Sanz-Garcia, Julio Fernandez-Ceniceros, Ruben 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.

13 Preprocesado de imágenes hiperespectrales en R

Rubén Urraca Valle, Borja Millán, Roberto Fernandez-Martinez, Andrés Sanz García

TELEVITIS. 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

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].

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

Roberto Fernandez Martinez, Ruben Lostado Lorza, Ruben 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.

15 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.

16 Estrategias de Captación de Clientes en Mercados con Competencia

Francisco Jesús Rodríguez Aragón
Doctor en Estadística por la Universidad de Córdoba
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 SABÍ 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

17 Sesgo de publicación en ciencias médicas

Borja Santos Zorrozuá1, 2, 3, Eduardo González Fraile4, , Javier Ballesteros Rodríguez2, 4

1 Universidad del País Vasco (UPV/EHU), 2 Cibersam (G16), 3 Programa PRE-DOC Gobierno Vasco, 4 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.

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

*Jose Antonio Palazon Ferrando y Antonio Maurandi López
Universidad de Murcia. Comunidad R-Hispano.
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’

Sesión de Comunicaciones III

Talleres

19 Relenium, selenium en R. Un nuevo paquete para webscraping.

Aleix Ruiz de Villa, Lluís Ramon, Andreu Vall

TSS - Transport Simulation Systems (<http://www.aimsun.com/wp/>)

RugBcn - grupo de usuarios de Barcelona (<http://rugbcn.wordpress.com/>)

Actualmente, los paquetes más utilizados para hacer web scraping con R són 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 language html y los xpath, y mostraremos las funcionalidades básicas del paquete reelenium. Lo complementaremos con las funcionalidades básicas de XML. No es necesario ningún conocimiento previo.

20 Cazando información espectro-temporal en datos ambientales con R

Dr. Josué M. Polanco Martínez

Investigador invitado, 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).

SOWAS: <http://tocsy.pik-potsdam.de/wavelets/> Biwavelet: <http://cran.r-project.org/web/packages/biwavelet/biwavelet.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 anexas 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).

Autores e Instituciones

Índice de autores

, 2, 21
, 3, 21
, 4, 21
, NA, 9, 21

Antonio Palazon Ferrando y Antonio Maurandi López, Jose, 22
Armenta, Ana, 9

Ballesteros Rodríguez², Javier, 21
Borja Santos Zorrozuá¹, , 21
Bouso Freijo, Jesús, 13
Breen, Martijn, 5

Casas, Jordi, 5
Concejero, Pedro, 9
Cordero, Jorge, 8
Couturier, Raphael, 15

de Blas Foix, Xavier, 19

Fernandez Martinez, Roberto, 18
Fernandez-Ceniceros, Julio, 16
Fernandez-Martinez, Roberto, 16, 17

García, César, 9
González Fraile⁴, Eduardo, 21
González Pérez, Beatriz, 8
Gregori, Pablo, 15
Gregorio Escalada, J., 9

Jesús Rodríguez Aragón, Francisco, 20
Jiménez Sanjuán, Sergio, 11
Josué M. Polanco Martínez, Dr., 14, 25

López López, Victoria, 8
Lostado Lorza, Ruben, 18
Luis Cañadas Reche, José, 2
Luis Ojeda Cabrera, Jorge, 4

Millán, Borja, 17

Pazmiño, Rubén, 15

Ramon, Lluís, 24
Ruiz de Villa, Aleix, 5, 24

Sanz García, Andrés, 17
Sanz Garcia, Andres, 18
Sanz-Garcia, Andres, 16

Torres-Manzanera, Emilio, 3

Urraca Valle, Rubén, 17
Urraca Valle, Ruben, 18
Urraca-Valle, Ruben, 16

Vall, Andreu, 24
Villegas, Paulo, 9

Índice de Instituciones

(Investigador invitado) Instituto de Economía Pública y Dept. , 14

1 Universidad del País Vasco (UPV/EHU), 2 Cibersam (G16), 3 Programa PREDOC Gobierno Vasco, 4 Instituto de Investigaciones Psiquiátricas, 21

Associate Professional Risk Manager, 20

Centro de Investigaciones Sociológicas (CIS), 13

Chronojump-Boscosystem., 19

de Econometría y Estadística, Universidad del País Vasco , 14

Departamento de Ecología e Hidrología. Facultad de Biología., 22

Department of Mining and Metallurgical Engineering and Materials Science. University of Basque Country, Spain, 16, 17

Department of Mining and Metallurgical Engineering and Materials Science. University of Basque Country. Spain, 18

Dept- Métodos Estadísticos, Univ. de Zaragoza, 4

Division of Bioscience. University of Helsinki, Finland, 16, 17

Division of Bioscience. University of Helsinki. Finland, 18

Doctor en Estadística por la Universidad de Córdoba, 20

EDMANS. Universidad de La Rioja, Spain, 16

Escuela Superior Politécnica de Chimborazo. Ecuador., 15

Investigador invitado, Instituto de Economía Pública y Dept. de Econometría y Estadística, Universidad del País Vasco. , 25

RugBcn - grupo de usuarios de Barcelona (<http://rugbcn.wordpress.com/>), 24

RugBcn - Grupo de usuarios de R de Barcelona, 5

SCIEN Analytics, 11

Sec. Apoyo Estadístico. Servicio de Apoyo a la Investigación (SAI), 22

Técnico de Investigación en el Instituto de Estudios Sociales Avanzados (IESA-CSIC), 2

Telefónica Digital, Product Development and Innovation, 9

TELEVITIS. Universidad de La Rioja, Spain, 16, 17

TSS - Transport Simulation Systems, 5

TSS - Transport Simulation Systems (<http://www.aim24.com>)

Universida Comte. Francia, 15

Universidad Complutense de Madrid, 8

Universidad de La Rioja. Spain, 18

Universidad de Murcia. Comunidad R-Hispano., 22

Universidad de Oviedo, 3

Universidad Jaume I. España., 15

Universitat Ramon Llull, FPCEE Blanquerna. Grupo SAFE., 19