*Actas*

Universidad Rey Juan Carlos
Servicio de Publicaciones

Carolina Cosculluela (Dir.)

# SEMINARIO INTERNACIONAL DE ESTADÍSTICA PARA LA INVESTIGACIÓN EN CIENCIAS SOCIALES. ENTORNO R.

ISBN 978-84-697-0439-4

# RELACIÓN DE PONENCIAS

## Contenido

RELACIÓN DE AUTORES

**Carolina Cosculluela Martínez**

**María del Carmen García-Centeno**

**LLuis Hurtado Gil**

**Raquel Ibar Alonso**

**Diego Mondéjar Ruiz**

**Javier Palencia González**

**Carlos Quesada González**

**Francisco Rabadán Pérez**

**Sonia Rodríguez-Sánchez**

**Alfonso Zamora Saiz**

# A VARMA MODEL COMPUTED IN R BY MEANS OF A ONE TERA HDD, NO RAM NEEDS

**Carolina Cosculluela Martínez**
**Departamento de Economía Aplicada I**
**Universidad Rey Juan Carlos**
**carolina.cosculluela@urjc.es**

**Raquel Ibar Alonso**
**Departamento Interfacultativo de Matemática Aplicada y Estadística**
**Universidad CEU San Pablo**
**ribar@ceu.es**

## Abstract

When using R program, which is an Opens Source program, the possibilities of processing Big Data is limited by the type of machine that you have or that you can rent for a short period, while the storage possibilities are higher and much cheaper. The entire instructions, libraries and packages needed for the estimation of the weights in the index proposed in *Time indicator of the Human Development Index* (Ibar-Alonso, Cosculluela-Martínez, & Hewings, 2017) published in Time and Society has been customized for the purpose. The goal is to increase the estimating speed whenever new variables and/or data from the same ones is required to improve the index, based on the library and package named filehash.

## Resumen

Cuando se usa el lenguaje de programación R, *Open Source*, las posibilidades de procesar Big Data están limitadas por el tipo de ordenador que se tiene o del que se puede disponer por un corto periodo de tiempo, mientras que las posibilidades de almacenamiento son más altas y mucho más económicas. Han sido programados los comandos, las librerías y los paquetes necesarios para la estimación de las ponderaciones en el índice propuesto en *Time indicator of the Human Development Index* (Ibar-Alonso, Cosculluela-Martínez, & Hewings, 2017) publicado en Time and Society. El objetivo incrementar la velocidad de estimación cuando se incluyan nuevas variables y/o se actualice la temporalización para mejorar el índice, utilizando la librería y el paquete *filehash*.

# PARADIGMA ECONÓMICO DEL BIG DATA

**Francisco Rabadán Pérez**
**Departamento de Economía Aplicada I**
**Universidad Rey Juan Carlos**
**francisco.rabadan@urjc.es**

## Abstract

Briefly discuss what is Big Data and how this has influenced the Business Intelligence. After that we will a classic analysis on the model of the 4 P's of Jerome McCarthy, adding a new dimension, trust. Finally, we will indicate a set of fundamental changes in the new economic paradigm that trigger strengths and weaknesses for the agents in the market.

## Resumen

Expondremos brevemente qué es el Big Data y como ha influido en la Business Intelligence. Posteriormente realizaremos un análisis clásico sobre el modelo de las 4 P's de Jerome McCarthy, añadiendo una nueva dimensión, la confianza. Finalmente indicaremos un conjunto de cambios fundamentales en el nuevo paradigma económico que desencadenan fortalezas y debilidades para los agentes en el mercado.

# PROCESOS PUNTUALES EN R.

**LLuis Hurtado Gil**
**Departamento Interfacultativo de Matemática Aplicada y Estadística**
**Universidad CEU San Pablo**
**lluis.hurtadogil@ceu.es**

## Abstract

Using maps in urbanism, agriculture, distribution of services and other works represents an important opportunity to study economics. Using Point Processes allows us to analyse with detail populations made of locations over a space, which incorporate quantitative and qualitative variables. This branch of statistics counts on abundant theoretical tools used before with success in biology (Pfeifer et al., 1992), industry (Berman, 1986), astrophysics (Peebles, 1980) or medicine (Diggle, 1990). In this work we propose an application on economics with the study of supermarkets distribution if Community of Madrid. Among the proposed statistical tools, we include descriptive methods (pair correlation function), inferential (hypothesis contrasts), data mining (nearest neighbour maps) and models of the underlying process (Gibbs models). The large amount of available data of this kind in economics requires of a great capacity of analysis that exceeds that of previously used methodologies, such as the Geographic Information Systems (GIS), this can be overcome using R libraries such as *spatstat*, designed for the analysis of Point Processes.

## Resumen

El uso de mapas en urbanismo, agricultura, distribución de servicios y otros trabajos representa una importante oportunidad para el estudio de la economía. El uso de Procesos Puntuales permite analizar con detalle poblaciones compuestas por localizaciones sobre un espacio en el que incorporan distintas variables de carácter cualitativo y cuantitativo. Esta rama de la estadística cuenta con un abundante desarrollo teórico que ya se ha usado con éxito en biología (Pfeifer et al., 1992), industria (Berman, 1986), astrofísica (Peebles, 1980) o medicina (Diggle, 1990). En este trabajo proponemos una aplicación en economía con el estudio de la distribución de supermercados en la Comunidad de Madrid. De entre las herramientas propuestas se incluyen métodos descriptivos (función de correlación de dos puntos), inferenciales (contrastes de hipótesis), minería de datos (mapas al vecino más próximo) y modelos del proceso subyacente (modelos de *Gibbs*). La gran cantidad de datos disponible de esta naturaleza en economía requiere de una mayor capacidad de estudio que excede la de metodologías anteriores como los *Geographic Information Systems* (GIS), esto puede resolverse con el uso de librerías de R como *spatstat*, diseñada para el análisis de Procesos Puntuales.

# ANÁLISIS TÉCNICO EN MERCADOS FINANCIEROS EN R

**Alfonso Zamora Saiz**
**Departamento Interfacultativo de Matemática Aplicada y Estadística**
**Universidad CEU San Pablo**
**Alfonso.zamorasaiz@ceu.es**

## Abstract

Technical Analysis stands for the identification of patterns in a time series of stock prices willing to forecast the future evolution. This disputes the Efficient-Market Hypothesis, understanding that market prices already contain all relevant information. Technical Analysis is based on using indicators constructed from price time series which generate buy and sell recommendations for the investor. In this talk we present some of the more common indicators, say Bollinger Bands, ADX, TDI and Parabolic SAR, their mathematical description, visualization over Banco Popular stock chart during 2016, R code implementation and the signals provided. Also, it is shown how the most popular financial servers, like Bloomberg or Investing, use these indicators to provide information for the trader.

## Resumen

El Análisis Técnico consiste en la identificación de patrones en una serie temporal de precios que intente predecir su evolución futura. Disputa la validez de la Hipótesis de Eficiencia de los Mercados al sostener que un precio precio ya refleja toda la información relevante. Se basa en la utilización de indicadores construidos a partir de la serie de precios que generan recomendaciones de compra y venta para el inversor.

En esta ponencia se presentan algunos de los Indicadores más comunes, como las Bandas de Bollinger, el ADX, el TDI y el SAR Parabólico, su descripción matemática, su visualización sobre la serie temporal de la acción del Banco Popular durante 2016, su implementación en código R y las señales que proporciona. También se muestra cómo los portales de finanzas más populares, como Bloomberg e Investing, usan estos indicadores para proporcionar información al inversor.

# LA TRANSICIÓN AL BIG DATA DE LA POLÍTICA DE PRECIOS DE LOS COMBUSTIBLES EN ESPAÑA

**Javier Palencia González**
**Departamento de Teoría Económica y Economía Matemática**
**Universidad Nacional de Educación a Distancia**
**jpalencia@cee.uned.es**

## Abstract

Historically the Spanish retail fuel market has been highly concentrated, and for a large part of the XXth century the prices were fixed administratively. In the mid-80s with a view to integration into the European Economic Community, EEC, the Spanish oil sector was deeply restructured for subsequent liberalization, which took place in the early 90s. From that moment the prices are fixed by the operators. The main sources for obtaining data on fuel prices are shown below. At European level, in 1999 there was a Council decision, then endorsed by another of the Commission, whereby all EU countries must remit every monday the prices of fuels in their territories. These data are compiled and published in the Weekly Oil Bulletin. In the case of Spanish, the prices are remitted by the administration and are an average of the prices published in the Peninsula and the Balearics Islands. At the same time the ministry creates a website www.geoportalgasolineras.es that allows make a query in real time. From the website you can download a daily file in xls format, which happens until May 2016. Since that moment you can download a file ever hour. Each file has the final prices of the different fuels marketed in all service stations in the country, currently more than 10,000. From another website of the Ministry you can get real-time files in different formats, JSON, XML, XLS and CSV. So we have moved from an administrative price to files with thousands of different prices in real time.

There are more sources that publish various reports and statistics related to the sector. The Strategic Reserves Corporation, CORES, publishes monthly statistics of Import and export of products by country of origin and destination, Consumption in regions and provinces and Production by refineries. The National Commission for Markets and Competition, CNMC, publishes monthly reports monitoring the fuel distribution in EE.SS, the CLH logistics activity and statistics of petroleum products. The Spanish Association of Operators of Petroleum Products, AOP, publishes an annual report that includes the state of the sector, consumption, product of the refineries, points of sale, etc. More data sources are the websites of companies, the press, the consulting and tax authorities, which publishes the collection of the Special Tax on Hydrocarbons, IEH. Finally we want to mention social networks and especially Twitter, where we can collect through programming in R all tweets with mentions a number of predefined terms like prices, stations and various brands.

## Resumen

Históricamente el mercado minorista español de combustibles ha estado muy concentrado, y durante una gran parte del S. XX los precios eran fijados administrativamente. A mediados de los años 80 con vistas a una integración en la Comunidad Económica Europea, CEE, el sector petrolero español fue profundamente reestructurado, para su posterior liberalización, lo que ocurre en los primeros años 90. A partir de ese momento los precios son fijados por los operadores del sector. Las principales fuentes que permiten obtener datos de los precios de los combustibles se muestran a continuación. A nivel europeo, en 1999 hubo una decisión del Consejo, luego refrendada por otra de la Comisión por la que todos los países de la UE deben remitir cada lunes los precios de los carburantes vigentes en sus territorios. Estos datos son recopilados y publicados en el Weekly Oil Bulletin con periodicidad semanal. En el caso español, los precios son remitidos por la administración y son una media de los precios publicados en península y Baleares. Al mismo tiempo el ministerio crea una web www.geoportalgasolineras.es que permite la consulta de datos en tiempo real. Desde la web se puede descargar un fichero diario en formato xls, lo que ocurre hasta mayo de 2016. Desde ese momento se puede descargar un fichero cada hora. Cada fichero tiene los precios finales de los distintos combustibles comercializados en todas las estaciones de servicio del país, actualmente más de 10.000. Desde la sede electrónica del Ministerio se pueden obtener ficheros en tiempo real en distintos formatos, JSON, XML, XLS y CSV. Por tanto se ha pasado de un precio administrativo a ficheros con miles de precios distintos en tiempo real.

Pero aún hay otras fuentes que publican diversos informes y estadísticas relativos al sector. La Corporación de Reservas Estratégicas, CORES, publica mensualmente estadísticas de Importación y exportación de productos por países de origen y destino, Consumo de CCAA y provincias y Producción por refinerías. La Comisión Nacional de los Mercados y la Competencia, publica el Informe mensual de supervisión de la distribución de carburantes en EE.SS., el Informe de supervisión de la actividad logística de CLH y la Estadística de productos petrolíferos. La Asociación Española de Operadores de Productos Petrolíferos, AOP, publica una memoria anual donde se recoge el estado del sector, consumo, producto de las refinerías, puntos de venta, etc.. Más fuentes de datos son las propias compañías mayoristas, la prensa, las consultoras del sector y la Agencia Tributaria, que publica la recaudación del Impuesto Especial de Hidrocarburos, IEH. Finalmente queremos mencionar las redes sociales y en particular twitter, donde podemos recabar a través de programación en R todos los tuits con menciones a una serie de términos prefijados como precios, estaciones y las diversas marcas.

# ESTIMACIÓN DE MODELOS DE VOLATILIDAD ESTOCASTICA. APLICACIÓN A SERIES DE RENDIMIENTOS DE METALES

**García-Centeno, María del Carmen**
**Departamento Interfacultativo de Matemática Aplicada y Estadística**
**Universidad CEU San Pablo**
**garcen@ceu.es**

**Rodríguez-Sánchez, Sonia**
**Departamento Interfacultativo de Matemática Aplicada y Estadística**
**Universidad CEU San Pablo**
**sonia.rodriguezsanchez@ceu.es**

## ABSTRACT

In financial literature, volatility is defined as a measure of the intensity in random and unpredictable changes in the return. The volatility need to be estimated because is not observable and it is used as a measure of risk in the financial market.

The models, that are often used to estimate it, are the conditional heterocedasticity models and the stochastic volatility models. The aim of this paper is to establish which type of model is the best to explain the behaviour of volatility in nine of daily yield of several types of metals. The results show that if the asymmetric behaviour of volatility exists (leverage effect), then the threshold asymmetric stochastic volatility model (TA-ARSV model) is the best model.

## RESUMEN

En la literatura financiera, la volatilidad se define como una medida de la intensidad de los cambios aleatorios e impredecibles en la rentabilidad. Como no es observable, debe ser estimada, ya que, en los mercados financieros se utiliza como una medida del riesgo.

Los modelos que se suelen utilizar para dicha estimación son: los modelos de heterocedasticidad condicional y los modelos de volatilidad estocástica. El objetivo de este trabajo es determinar qué tipo de modelo es más adecuado para explicar el comportamiento de la volatilidad en nueve series de rendimientos diarios de diferentes tipos de metales. Los resultados muestran que si existe respuesta asimétrica de la volatilidad (efecto leverage), entonces es mejor el modelo de volatilidad estocástica asimétrica por umbrales (modelo TA-ARSV).

# LA FORMA DE LOS DATOS. UNA PERSPECTIVA GEOMÉTRICA DEL BIG DATA.

**Diego Mondéjar Ruiz**
**Departamento Interfacultativo de Matemática Aplicada y Estadística**
**Universidad CEU San Pablo**
**Diego.mondejarruiz@ceu.es**

## ABSTRACT

The Topological Data Analysis (TDA) is a relatively new discipline in which we use techniques and results of topology to find topological or geometrical structure in massive data. These structures are able to discover patterns and qualitative characteristics of a point cloud. In the last years there is a huge increase of the use of these techniques and the results obtained are surprising. Furthermore, many companies are starting to offer TDA based analysis of data to offer important and profitable information to its customers.

## RESUMEN

El análisis topológico de datos es una disciplina relativamente nueva en la que, utilizando resultados y técnicas propias de la topología, se trata de encontrar estructuras geométricas o topológicas en grandes cantidades de datos. Estas estructuras pueden revelarnos comportamientos, patrones y características cualitativas ocultas en la nube de puntos que se analiza. En los últimos años, ha tenido lugar un gran aumento del uso de estas técnicas dando lugar a trabajos con resultados sorprendentes. Por otro lado, muchas empresas están comenzando a utilizar estas herramientas para analizar datos de cualquier procedencia para obtener información que pueda resultar de utilidad para sus clientes.

# WEB SCRAPING EN R.

**Carlos Quesada González**
**Departamento Interfacultativo de Matemática Aplicada y Estadística**
**Universidad CEU San Pablo**
**carlos.quesadagonzalez@ceu.es**

## ABSTRACT

Obtaining data is the first step in statistical learning. Sometimes we are lucky enough to have a well formatted database from where we can retrieve data. But many times the information, although available on the internet, is not stored in a database but as scattered text in a web, or even many webs. In order to automatically get that information and properly save it into a format that we can work with, we need *web scraping*.

We introduce the concept of web scraping, showing how to analyze the html code of a web through SelectorGadget to spot the html nodes we are interested in. Then, we use the package *rvest* in R to explore that nodes in the web and automatically retrieve the data and store it in a data table in an automatic way. Finally, the limitations of *rvest* are explored in order to introduce the need of *Rselenium* for complicated cases.

## RESUMEN

La obtención de datos es el primer paso para el aprendizaje estadístico. A veces tenemos la suerte de contar con una base de datos ordenada de donde recoger los datos. Sin embargo, muchas veces, y a pesar de estar disponible en internet, la información no está organizada ni almacenada en bases de datos, sino que está dispersa en una web, o incluso varias. Para conseguir toda esta información automáticamente y almacenarla en un formato con el que poder trabajar necesitamos el *web scraping*.

Introducimos este concepto, mostrando cómo analizar el código html de una web utilizando la herramienta SelectorGadget para encontrar los nodos html en los que estamos interesados. Posteriormente, usamos el paquete *rvest*de R para explorar dichos nodos y recoger y almacenar la información en *data tables* de forma automática. Finalmente, se exploran las limitaciones de *rvest* para introducir la necesidad de *Rselenium* en los casos más complicados.