KNITR Y RMARKDOWN: MODELOS TEXTOS AUTO-ALIMENTADOS Y AUTO-EXPLICADOS

Francisco J. Rodríguez Aragón

Dr. En Estadística (Univ. Córdoba)

Responsable de Análisis Estadístico en CESCE Consulting

INTRODUCCIÓN

- El trabajo que se presenta aquí aún está en preparación y se esperan nuevos resultados más adelante
- Aquí se ofrecen las ideas principales que permite vincular, análisis estadístico, presentación e interpretación de resultados y automatización de todo esto
- Todo el código desarrollado por el momento es totalmente reproducible en cualquier ordenador con conexión a internet y con las librerías R adecuadas
- Por el momento se hace uso sólo de datos que el Banco de España publica en su Boletín Estadístico

CONCEPTO DE INVESTIGACIÓN REPRODUCIBLE

- El 18 de Febrero del 2015, un día antes de mi cumpleaños, se publica en:

http://www.investigacionyciencia.es/blogs/medicina-y-biologia/69/posts/es-reproducible-la-ciencia-que-se-publica-12880

- Se hacen afirmaciones como:

  - Que casi la mitad de los trabajos en biomedicina no resultan reproducibles por otros investigadores
  - Que los científicos de la farmaceútica Bayer no pudieron reproducir el 75% de los trabajos realizados sobre enfermedades cardio-vasculares

CONCEPTO DE INVESTIGACIÓN REPRODUCIBLE

- Una de las primeras críticas aunque también en el campo de la medicina se encuentra en (y data del 2006):

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/

- Pero no sólo en campos de la medicina existe este problema, en casos aún más sencillos donde no hay que seguir los pasos exactos de un experimento de laboratorio en ocasiones ocurre:

  - Que tras haber publicado un trabajo alguien pide las fuentes originales y éstas no existen, aunque sea sólo para consultarlas
  - No se tienen referencias exactas a éstas y por tanto, si se actualizan los datos es imposible actualizar las conclusiones
  - Etc

CONCEPTO DE INVESTIGACIÓN REPRODUCIBLE

- En R ya se ha tratado este problema desde hace tiempo, primero con sweave como se puede observar el siguiente post de Datanalytics del 2011:

http://www.datanalytics.com/2011/06/23/sweave-investigacion-reproducible-y-mas/

- La demo que presentaremos aquí, dentro del entorno Rmarkdown, cumple con estos estándares:

> The term *reproducible research* refers to the idea that the ultimate product of academic research is the paper along with the full computational environment used to produce the results in the paper such as the code, data, etc. that can be used to reproduce the results and create new work based on the research

TEXTOS AUTO-ALIMENTADOS

- Un texto podría considerarse como Auto-Alimentado, cuando es capaz de acceder a las fuentes originales de los datos que ofrece y actualizarlos acorde a dichas fuentes:
  - Este tipo de textos pueden estar permanentemente conectados a las fuentes de información de los que se nutren
  - En estos textos o informes, cambian los valores de los datos que muestran cuando éstos cambian

TEXTOS AUTO-EXPLICADOS

- Un texto podría considerarse como Auto-Explicado, cuando no es necesaria la intervención humana para la actualización de su contenido acorde a los cambios de los datos en los que éstos se basan

  - Este tipo de textos pueden estar permanentemente conectados a las fuentes de información de los que se nutren
  - Además tienen bases de datos particulares con reglas y conclusiones en función del valor de los datos que en todo caso consideran
  - A lo sumo la única función humana que debería poderse considerar es la obligación a actualizar la información en el instante actual, aunque en casos más avanzados, el mismo texto puede informar cuando hay alguna variación en sus datos más o menos significativa

RMARKDOWN Y KNITR

- A continuación se desarrolla un ejemplo sencillo con las anteriores ideas bajo entorno RMarkdown, antes algunas indicaciones:
  - Instalar Rmarkdown y que reproduzca textos en .html, es bastante sencillo y rápido
  - Para que los textos se reproduzcan en formato .pdf, como aquí se hace, es necesario programas adicionales y la instalación es bastante larga. Al final se consigue generar documentos .pdf en formato tipo LATEX
  - En cualquiera de los dos casos (html o pdf), este estudio, con el código que se dejará colgado es totalmente reproducible con tal de tocar el botón “knit html” o “knit pdf” que se comenta a continuación
  - Cuando se usa knitr para reproducir las plantillas, análogos cambios serán considerados

RMARKDOWN Y KNITR

- Fuente de datos:

www.bde.es/f/webbde/SES/Secciones/Publicaciones/InformesBoletinesRevistas/BoletinEstadistico/15/Fich/bes1509.zip

- Código R de conexión a datos:

```
#Interfaz preliminar

anio <- as.character("15") #Año 2015
mes <- as.character("08")  #Mes de Agosto
```

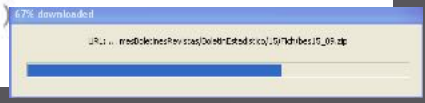
Se elige el año y la fecha del fichero de datos a acceder (cuando se realizó este código, aún no estaba el boletín de Septiembre del 2015)

```
ser1 <- "BE_1_1.1"
ser2 <- "BE_1_1.2"
ser3 <- "BE_1_1.3"
ser5 <- "BE_1_1.5"
ser6 <- "BE_1_1.6"
ser8 <- "BE_1_1.8" #Contabilidad Nacional/METODOLOGÍA DEL AÑO SEC2010/
                   #Año base 2010/Volúmenes encadenados/Producto interior bruto/
                   #Empleos. Economía en su conjunto (Total de la economía)/Recursos. Economía en su
                   #/Zona del Euro/Datos corregidos de efectos estacionales y de calendario
```

Nombre de columnas del fichero base del BdE

Código R de bajada de datos. Al estar los datos en un fichero comprimido, primero se baja el .zip y después se elijen los .csv que se desean leer: *catalogo_be.csv* y *be0101.csv*

```
temp <- tempfile()
download.file(paste0("http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/
                     InformesBoletinesRevistas/BoletinEstadistico/",anio,"/Fich/
                     bes",anio,"_",mes,".zip"),temp)
cat_series <- read.csv(unz(temp, "catalogo_be.csv"))
serie_sel <- read.csv(unz(temp, "be0101.csv"))
unlink(temp)
```



RMARKDOWN Y KNITR

- Código R de depuración de datos:

```r
serie_datos_euro <- function (serie){

  #Se llama a la serie de datos en sí
  serie_sel_char <- as.character(get(serie))

  #Se normalizan los datos
  serie_sel_char <- ifelse(serie_sel_char == "_", "",serie_sel_char)
  serie_sel_char[1] = ""
  serie_sel_char[2] = ""
  serie_sel_char[3] = ""

  serie_sel_num <- as.numeric(serie_sel_char)
  serie_sel_num1 <- na.omit(as.numeric(serie_sel_char))

  #Se convierten los datos a serie temporal
  return (ts(serie_sel_num1, frequency = 4, start = c(1995, 1)))

}
```

Se construye una función adaptada a las características de las series a leer para la adecuada preparación de los datos

El formato base que ofrecen los datos del BdE son poco tratables por R tal y como se observa en esta muestra

	A	B	C	D
1		BE_1_1.1	BE_1_1.2	BE_1_1.3
2	NÚMERO SEC	2665105	2665106	2665107
3	DESCRIPCIÓN	Contabilidad	Contabilidad	Contabilidad
4	DESCRIPCIÓN	Millones de	Millones de	Millones de
5	ENE 1995	_	_	_
6	Feb-95	_	_	_
7	Mar-95	1002807.71	367786.89	370388.03
8	ABR 1995	_	_	_
9	May-95	_	_	_
10	Jun-95	1014149.1	371165.03	374272.05
11	Jul-95	_	_	_
12	AGO 1995	_	_	_
13	Sep-95	1013114.91	373512.28	372925.34

```r
serie1 <- serie_datos_euro(ser1)
```

	Qtr1	Qtr2	Qtr3	Qtr4
1995	1002808	1014149	1013115	1014699
1996	1022222	1025913	1031679	1032456
1997	1035820	1044580	1046695	1059063
1998	1065738	1073102	1082352	1093034
1999	1099785	1107300	1118625	1128314
2000	1137563	1147263	1151715	1153647
2001	1197659	1202100	1206139	1205143
2002	1206391	1208972	1215163	1221279
2003	1220110	1222438	1227938	1230749
2004	1238185	1241507	1244100	1254654
2005	1257851	1264367	1271789	1277106
2006	1284080	1290670	1294008	1305963
2007	1310344	1318256	1323044	1327807
2008	1333206	1330299	1323968	1319857
2009	1324889	1324435	1321977	1327730
2010	1329101	1333460	1335069	1340505
2011	1341090	1335378	1336738	1328929
2012	1326119	1320731	1318949	1311085
2013	1307191	1309634	1312843	1314698
2014	1319363	1322089	1328752	1336004
2015	1347658	1352595		

RMARKDOWN Y KNITR

- Fichero Rmarkdown: Texto + Código

```
---
title: "Contabilidad Nacional de la Zona Euro"
output: html_document
---

Se presenta la variaciones trimestrales interanuales del PIB de la zona Euro junto con las variaciones trimestrales
interanuales de cada uno de sus componentes desde el punto de vista de la demanda:

<br></br>

$$PIB = Consumo + Gasto + FBCF + VE + (Exportaciones - Importaciones)$$

<br></br>

  * **Consumo:** consumo privado por pate de los hogares de las Instituciones sin Fines de Lucro al Servicio de los Hogares
(ISLSH)

  * **Gasto:** gasto final de las Administraciones Públicas (AAPP)

  * **FBCF:** Formación Bruta del Capital Fijo, constituida por los Activos fijos materiales y los inmateriales

  * **VE:** Variación de Existencias. Esta parte no se considera a este nivel agregado de análisis

  * **Exportaciones:** Exportaciones de bienes, servicios y consumo de no residentes en territorio económico

  * **Importaciones:** Importaciones de bienes, servicios y consumo de residentes en el resto del mundo

```{r, echo = FALSE}

anio <- as.character("15") #Año 2015
mes <- as.character("08")  #Mes de Agosto

ser1 <- "BE_1_1.1"
ser2 <- "BE_1_1.2"
ser3 <- "BE_1_1.3"
ser5 <- "BE_1_1.5"
```

```
ser6 <- "BE_1_1.6"
ser8 <- "BE_1_1.8" #Contabilidad Nacional/METODOLOGÍA DEL AÑO SEC2010/
                   #Año base 2010/Volúmenes encadenados/Producto interior bruto/
                   #Empleos. Economía en su conjunto (Total de la economía)/Recursos. Economía en su conjunto (Total de la
economía)/Euro área (composición variable)
                   #/Zona del Euro/Datos corregidos de efectos estacionales y de calendario


temp <- tempfile()
download.file(paste0("http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/InformesBoletinesRevistas/BoletinEstadistico/",
anio,"/Fich/bes",anio,"_",mes,".zip"),temp)
cat_series <- read.csv(unz(temp, "catalogo_be.csv"))
serie_sel <- read.csv(unz(temp, "be0101.csv"))
unlink(temp)

attach(serie_sel)

serie_datos_euro <- function (serie){

  #Se llama a la serie de datos en sí
  serie_sel_char <- as.character(get(serie))
```

El texto y el código están en un mismo documento .Rmd que genera un PDF cada vez que se ejecuta
```

RMARKDOWN Y KNITR

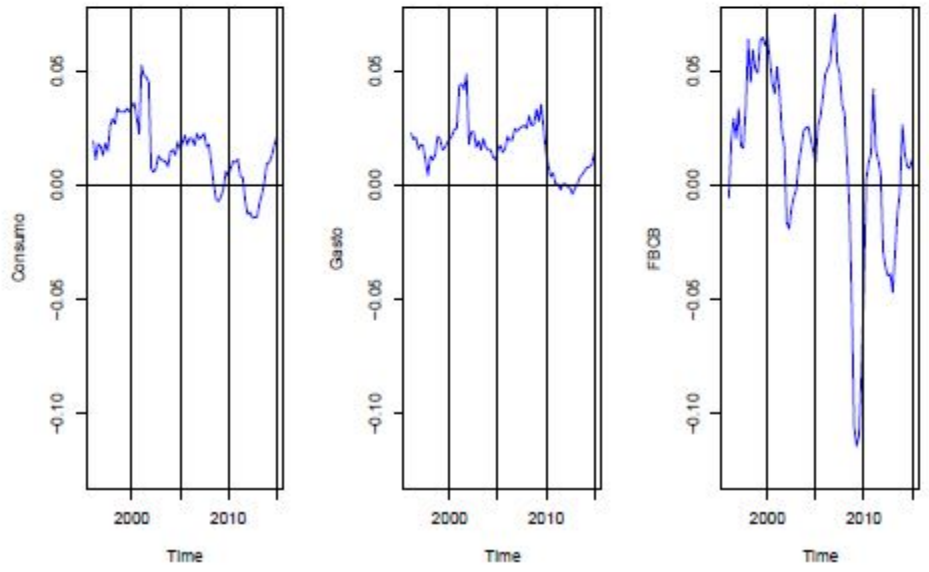
- Resultado final:

Contabilidad Nacional de la Zona Euro

Se presentan la variaciones trimestrales interanuales del PIB de la zona Euro junto con las variaciones trimestrales interanuales de cada uno de sus componentes desde el punto de vista de la demanda:

$$PIB = Consumo + Gasto + FBCF + VE + (Exportaciones - Importaciones)$$

- **Consumo:** consumo privado por pate de los hogares de las Instituciones sin Fines de Lucro al Servicio de los Hogares (ISLSH)
- **Gasto:** gasto final de las Administraciones Públicas (AAPP)
- **FBCF:** Formación Bruta del Capital Fijo, constituida por los Activos fijos materiales y los inmateriales
- **VE:** Variación de Existencias. Esta parte no se considera a este nivel agregado de análisis
- **Exportaciones:** Exportaciones de bienes, servicios y consumo de no residentes en territorio económico
- **Importaciones:** Importaciones de bienes, servicios y consumo de residentes en el resto del mundo



RMARKDOWN Y KNITR

- Generalización: render()

```
library(rmarkdown)

fechas <- c("082015","092015")

for (i in fechas){

  write.csv(i, "C:\\FRANCISCO\\2015\\BDE\\DATOS\\Configuracion.csv")

  render(input = "C:\\FRANCISCO\\2015\\BDE\\PROYECTOS\\INDICADORES_01.Rmd",
         output_file = paste0("C:\\FRANCISCO\\2015\\BDE\\PROYECTOS\\SALIDA_FINAL\\INDICADORES_",
                              i, ".pdf"))
}
```

Permite considerar el documento como una plantilla donde cambiar los datos conforme se cambia el tiempo

Se define un fichero de un único dato que permite introducir los cambios en RMarkdown

render() permite ejecutar desde R (y en modo batch) la plantilla creada y enviar los resultados a ficheros diferenciados

RMARKDOWN Y KNITR

- El fichero Rmd, se adapta para que pueda funcionar como una plantilla con una ligera variación en su código (re-definición de las variables *mes* y *anio*)

El fichero Configuracion.csv es modificado por el código R anterior y es llamado por Rmarkdown

```
```{r, echo = FALSE, message = FALSE}

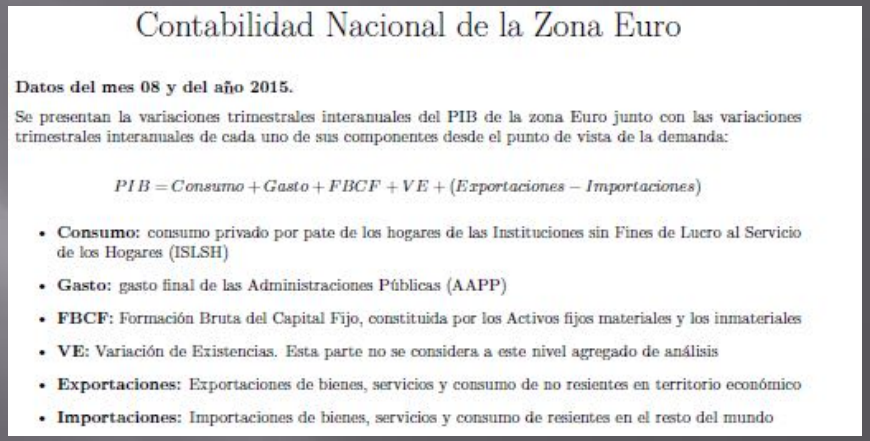
CONF <-read.csv("C:\\FRANCISCO\\2015\\BDE\\DATOS\\Configuracion.csv", sep = ",")
FECHA <- as.character(CONF$x)
FECHA <-ifelse(nchar(FECHA)< 6, paste0("0",FECHA) , FECHA)

mes <- substr(FECHA, 1, 2)
anio <- substr(FECHA, 5, 6)
```
```

La variables mes y anio dependen de los datos que existen en Configuracion.csv

RMARKDOWN Y KNITR

- Resultado final tras aplicar una opción bucle simple



Contabilidad Nacional de la Zona Euro

Datos del mes 08 y del año 2015.

Se presentan la variaciones trimestrales interanuales del PIB de la zona Euro junto con las variaciones trimestrales interanuales de cada uno de sus componentes desde el punto de vista de la demanda:

$$PIB = Consumo + Gasto + FBCF + VE + (Exportaciones - Importaciones)$$

- **Consumo:** consumo privado por pate de los hogares de las Instituciones sin Fines de Lucro al Servicio de los Hogares (ISLSH)
- **Gasto:** gasto final de las Administraciones Públicas (AAPP)
- **FBCF:** Formación Bruta del Capital Fijo, constituida por los Activos fijos materiales y los inmateriales
- **VE:** Variación de Existencias. Esta parte no se considera a este nivel agregado de análisis
- **Exportaciones:** Exportaciones de bienes, servicios y consumo de no residentes en territorio económico
- **Importaciones:** Importaciones de bienes, servicios y consumo de residentes en el resto del mundo

- En este caso se está ante un análisis muy sencillo, pero tal y como puede observarse en el código, sería posible incorporar modelizaciones más avanzadas para poder predecir comportamientos a futuros de series temporales en función de la información disponible

CONCLUSIONES FINALES

- Rmarkdown permite la integración de acceso a datos, cálculo estadístico-matemático, texto, tablas, imágenes, etc; en un mismo documento

- Rmarkdown ofrece en un formato bastante legible los resultados de una determinada investigación, permitiendo reproducir todos los pasos que van desde el acceso a los datos hasta la presentación final de resultados y su interpretación

- Rmarkdown permite cálculo iterativo de trabajos de investigación y de sus resultados

CONCLUSIONES FINALES

- Es posible programar códigos que dependen de fechas y de tiempos para que se ejecuten automáticamente

- Lo anterior permitiría la incorporación de modelos que sean capaces de realizar, predicciones, interpretarlas, ofrecer conclusiones y cambiar las anteriores conforme llegue nueva información con poca o ninguna intervención humana