Descarga de datos del Instituto Nacional de Estadística con R usando el servicio API JSON

Daniel Redondo-Sánchez, Miguel Ángel Luque Fernández, Miguel Rodríguez Barranco, Pablo Fernández-Navarro, María José Sánchez Pérez

Instituto de Investigación Biosanitaria de Granada (ibs.GRANADA), Universidad de Granada

Registro de Cáncer de Granada, Escuela Andaluza de Salud Pública

CIBER de Epidemiología y Salud Pública

"Una manera de hacer Europa"

Objetivo

Describir un **método de descarga de información** del Instituto Nacional de Estadística (INE) usando su servicio API y el software R.

El código es totalmente reproducible y está disponible en GitHub:

github.com/danielredondo/INE_R

Utilización del servicio API del INE

Utilizamos el servicio API (*Application Programming Interface*) del INE para realizar la tarea de conexión e intercambio de datos.

1. **Obtenemos la dirección web** válida para la descarga, en función del tipo de información a descargar (por ejemplo, si es una tabla con número determinado, o si es un fichero *PCAxis*).

2. Después, procedemos a la **descarga de información** usando el comando GET del paquete **httr** (v1.4.0). El contenido se descarga en formato JSON (*JavaScript Object Notation*).

3. Se **procesa la información** con **dplyr** (v0.8.3), **rlist** (v0.4.6.1) y **data.table** (v1.12.4) hasta obtener un objeto data.frame para su fácil manipulación en R.

1. Obtención de URL


```
url <- "http://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/ . . . "
```

1. Obtención de URL

Tipo 1: Número de tabla

URL descarga:
http://ine.es/jaxiT3/Tabla.htm?t=9687

URL API:
http://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/9687

Tipo 2: Fichero PC-AXIS

URL descarga:
http://ine.es/jaxi/Tabla.htm?path=/t15/p417/a2017/l0/&file=01007.px

URL API:
http://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/t15/p417/a2017/l0/01007.px

2. Descarga de información


```
library(httr)
datos_json <- GET(url)
datos_json$status_code
```

Si empieza por:	Suele indicar:
2 ó 3	Éxito
4	Error (del código)
5	Error (de la web)

3. Procesamiento de datos


```
datos <- content(datos_json)
```

+

```
%>%
rlist::list.select
rlist::list.stack
data.table::rbindlist
```

= data.frame

Ejemplo: procesamiento de defunciones fetales tardías

```
n <- length(defunciones_contenido)

for(i in 1:n){
   dato.i <- defunciones_contenido[[i]]
   defunciones.i <- list.select(dato.i$Data, Valor) %>% list.stack %>%
         cbind(codigo = dato.i$MetaData[[1]]$Codigo) %>%
         cbind(sexo = dato.i$MetaData[[2]]$Codigo) %>%
         cbind(s_gest = dato.i$MetaData[[3]]$Codigo)
   ifelse(i == 1,
          defunciones <- defunciones.i,
          defunciones <- rbindlist(list(defunciones, defunciones.i)))
}

head(defunciones)
```

```
##    Valor                  codigo       sexo           s_gest
## 1:  1274 0193ixxiitodaslascausas ambossexos            total
## 2:   196 0193ixxiitodaslascausas ambossexos menosde28semanas
## 3:   229 0193ixxiitodaslascausas ambossexos   de28a31semanas
## 4:   355 0193ixxiitodaslascausas ambossexos   de32a36semanas
## 5:   322 0193ixxiitodaslascausas ambossexos   de37a41semanas
## 6:     0 0193ixxiitodaslascausas ambossexos    42semanasymas
```

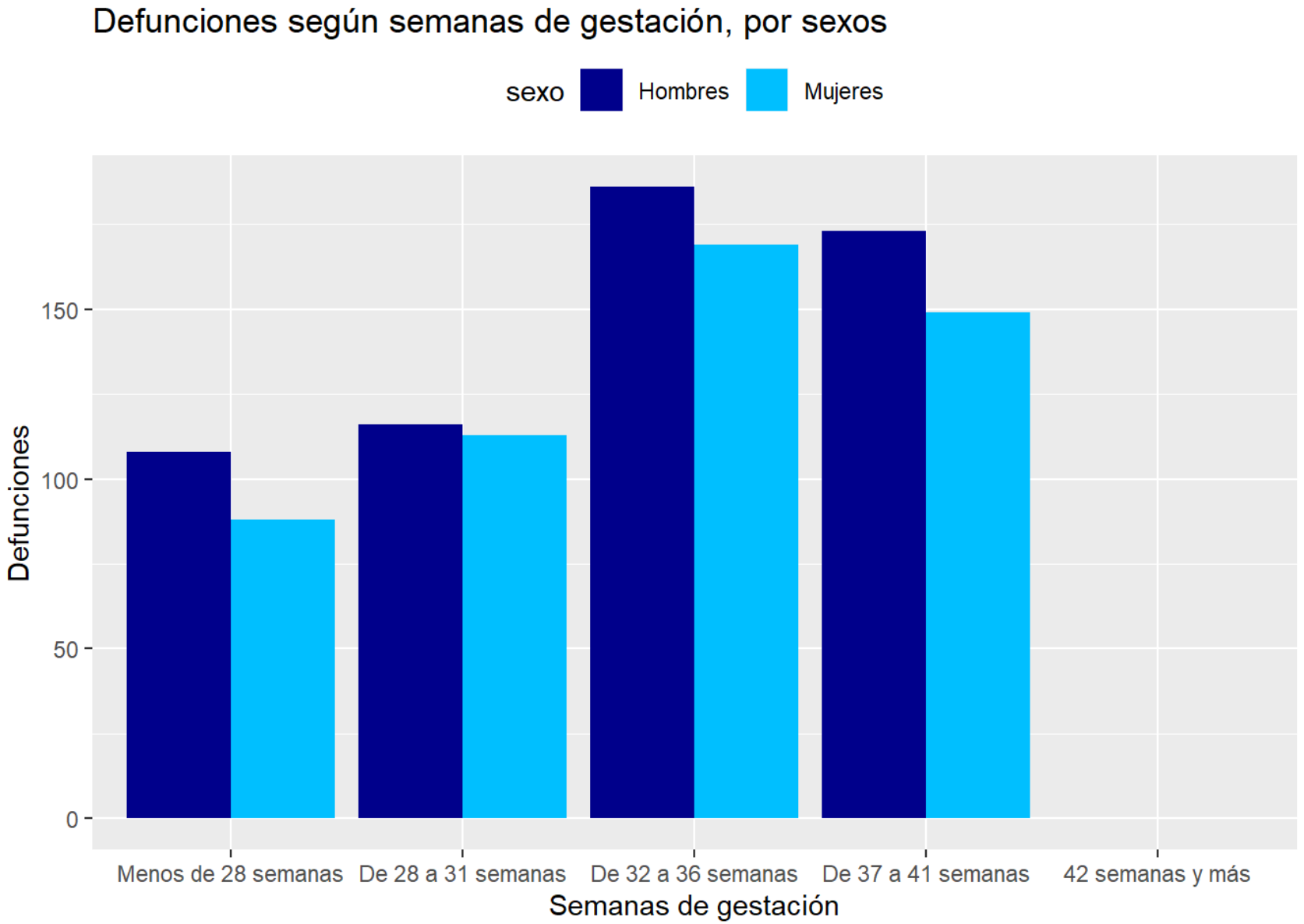
Aplicabilidad de uso de los datos descargados

```r
library(ggplot2)

datos <- subset(defunciones, defunciones$codigo == "01-93 I-XXII.Todas las causas" &
                  defunciones$sexo != "Ambos sexos" &
                  ! defunciones$s_gest %in% c("Total", "No consta"))

ggplot(datos, aes(x = s_gest, y = Valor, fill = sexo)) +
  scale_fill_manual(values=c("darkblue", "deepskyblue")) +
  geom_bar(stat = "identity",  position = "dodge") +
  ylab("Defunciones") + xlab("Semanas de gestación") +
  ggtitle("Defunciones según semanas de gestación, por sexos") +
  theme(legend.position="top")
```

Aplicabilidad de uso de los datos descargados



Fortalezas y limitaciones

Fortaleza: **Fácil implementación**

En GitHub (github.com/danielredondo/INE_R) está disponible un **tutorial** (Rmd/html) con dos ejemplos de descarga:

- Defunciones fetales tardías por causas (lista perinatal), sexo y semanas de gestación, 2017.
- Población por provincias, por edad simple, 2017-2019.

Fortaleza: **Reproducibilidad**

Esta descarga permite **reproducibilidad** en análisis posteriores, algo importante para lograr **transparencia** en la publicación de resultados científicos.

Fortalezas y limitaciones

Fortaleza: **Descarga masiva de información**

Haciendo uso de **156 URLs diferentes** (52 provincias $\times$ 3 años) realizamos una descarga automática de **+6.000.000 de filas** (población por edad simple por secciones censales de los años 2010, 2011 y 2012).

```
[1] "Se han descargado 17556 filas de Araba/Álava"
[1] "Se ha anexado la información de Araba/Álava"
[1] "Se han descargado 19866 filas de Albacete"
[1] "Se ha anexado la información de Albacete"
[1] "Se han descargado 80982 filas de Alicante/Alacant"
[1] "Se ha anexado la información de Alicante/Alacant"
[1] "Se han descargado 27984 filas de Almería"
[1] "Se ha anexado la información de Almería"
[1] "Se han descargado 20724 filas de Avila"
[1] "Se ha anexado la información de Avila"
[1] "Se han descargado 36498 filas de Badajoz"
[1] "Se ha anexado la información de Badajoz"
```

Limitación: **Tiempo**

Largos tiempos de espera en la descarga y procesamiento de información. Recomendable el uso de RStudio con **Google Cloud** o **Amazon Web Services** (AWS).

¡Gracias!

Daniel Redondo Sánchez

✉ daniel.redondo.easp@juntadeandalucia.es
danielredondo.com
github.com/danielredondo
@dredondosanchez

Financiación: Instituto de Salud Carlos III (FIS PI18/01593 EU-FEDER)
Subprograma de Vigilancia Epidemiológica del Cáncer del CIBERESP

ibs.GRANADA
INSTITUTO DE INVESTIGACIÓN BIOSANITARIA

registro de cáncer de granada

VICA

JUNTA DE ANDALUCIA
Escuela Andaluza de Salud Pública
CONSEJERÍA DE SALUD Y FAMILIAS

ciberesp

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Instituto de Salud Carlos III

UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional
"Una manera de hacer Europa"