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

Daniel Redondo Sánchez (Instituto de Investigación Biosanitaria ibs.GRANADA, Escuela Andaluza de Salud Pública, CIBERESP); **Miguel Ángel Luque Fernández** (Instituto de Investigación Biosanitaria ibs.GRANADA, Escuela Andaluza de Salud Pública, CIBERESP), **Miguel Rodríguez Barranco** (Instituto de Investigación Biosanitaria ibs.GRANADA, Escuela Andaluza de Salud Pública, CIBERESP), **Pablo Fernández-Navarro** (Centro Nacional de Epidemiología, CIBERESP), **María José Sánchez Pérez** (Instituto de Investigación Biosanitaria ibs.GRANADA, Escuela Andaluza de Salud Pública, CIBERESP).

En este trabajo describimos un método de descarga de información del Instituto Nacional de Estadística (INE) usando R. El código es totalmente reproducible y está disponible en un repositorio de GitHub: https://github.com/danielredondo/INE_R

Utilizamos el servicio API (Application Programming Interface) del INE para realizar la tarea de conexión e intercambio de datos. En primer lugar, obtenemos la URL (dirección web) válida para la descarga, en función del tipo de información a descargar (por ejemplo, si es una tabla con número definido, o si es un fichero PCAxis). 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), y es posteriormente procesado con dplyr (v0.8.3) y rlist (v0.4.6.1) hasta obtener un objeto data.frame para su fácil manipulación en R.

Mostramos dos ejemplos donde descargamos datos de las estadísticas vitales de defunciones perinatales precoces y tardías según semanas de gestación, y las cifras de población en España más recientes, por edad y provincia. Además, acompañamos la descarga de la información con gráficos realizados con ggplot2 (v3.2.0) para facilitar la interpretación y visualización de la información descargada.

Finalmente, mostramos la utilidad de la aplicación para la descarga de grandes volúmenes de información (Big Data), realizando descarga automática de más de seis millones de filas (población por edad simple por secciones censales de los años 2010, 2011 y 2012), haciendo uso de 156 URLs diferentes (52 provincias, 3 años). El tiempo aproximado de descarga fue de 75 horas en un ordenador con 8Gb de RAM.

Financiación: Instituto de Salud Carlos III (FIS PI18/01593 EU-FEDER), Subprograma de Vigilancia Epidemiológica del Cáncer (VICA) del CIBER de Epidemiología y Salud Pública (CIBERESP).