TRATAMIENTO DE GRANDES VOLÚMENES DE DATOS CON TIDYVERSE. EJEMPLOS A PARTIR DE LA MCVL

Isabel Padilla Sánchez
Instituto de Estadística y Cartografía de Andalucía
isabel.padilla@juntadeandalucia.es

Joaquín Planelles Romero
Instituto de Estadística y Cartografía de Andalucía
joaquin.planelles@juntadeandalucia.es

PALABRAS CLAVE
R, tidyverse, dbplyr, MCVL, trayectorias laborales

RESUMEN

En la última década la demanda y la oferta de datos se han expandido enormemente. Un reflejo de este proceso lo encontramos en el desarrollo de diversos paquetes informáticos, tanto de naturaleza libre como comercial, que facilitan el tratamiento de grandes volúmenes de datos. En este artículo nos vamos a centrar en algunas librerías de R específicas para estos fines.

El tratamiento de las bases de datos presenta distintos tipos de dificultades. Por un lado, el almacenamiento. Se necesitan máquinas potentes que almacenen la información. Otro aspecto importante a tener en cuenta es eficiencia en el tratamiento de los datos, es decir, al trabajar con grandes volúmenes de datos los tiempos de respuesta son mayores, lo que dificulta el trabajo. Esto pone de manifiesto la necesidad de utilizar programas eficientes y que el personal implicado en el tratamiento de los datos esté suficientemente cualificado para ello.

R es uno de los lenguajes de programación más extendidos a nivel mundial. Además, cuenta con una extensa comunidad de desarrolladores, que han ido creando librerías para problemas concretos (en el repositorio oficial de CRAN hay más de 10 mil librerías). Esta ponencia versa sobre un conjunto de paquetes de R que resuelven parte de los problemas que surgen en el tratamiento de datos, utilizando para ello una sintaxis a la vez sencilla de entender y eficiente, el entorno 'tidyverse'.

No obstante, R presenta ciertas limitaciones al trabajar con grandes volúmenes de datos, ya que consume recursos de la memoria volátil del ordenador. Existen diversas estrategias para resolver estas limitaciones. En esta ponencia se propondrá una alternativa posible, mediante el uso de la librería dbplyr, perteneciente también al entorno tidyverse. Esta librería funciona como un traductor a SQL. De este modo, es posible alojar la información en una base de datos externa y conectarse a la misma desde R.

En la ponencia se mostrará esta filosofía de trabajo mediante el trabajo que se ha desarrollado en el IECA con la Muestra Continua de Vidas Laborales. Se trata de un conjunto voluminoso de datos almacenados en una base de datos mediante tablas relacionales a la que se accede desde R y sobre los que se extrae información resumida mediante las librerías de tidyverse.