ID de aportación : **27** Tipo: **Taller**

Uso de R en la producción estadística: caso de uso en el IECA

En cualquier trabajo con datos hay distintas fases. Típicamente el proceso comienza con la lectura/captación de la información. Y esto puede implicar acciones muy diversas. En unos casos serás tú mismo el que capte esa información (preguntando, observando, sensorizando) y en otras ocasiones -las más- reutilizarás información que ha sido captada por terceros con fines distintos a la producción de estadísticas o modelización de los datos. Por ejemplo, información que ha ido quedando sedimentada en una base de datos para la prestación de un servicio. En otras ocasiones acudirás a información menos estructurada, haciendo Web-scrapping o tratando imágenes (aquí la escala va de los satélites a los cultivos microscópicos en laboratorio).

Una vez has captado la información, el siguiente paso es la transformación de la estructura y contenido de tu set de datos. Esta segunda fase, dedicada al tratamiento de datos, es la "fase gris"del ciclo de vida del dato. Le falta el atractivo visual que tienen las fases siguientes, centradas en la modelización y la comunicación. No obstante, por experiencia sabemos que una parte significativa del tiempo total de trabajo se dedica a este tipo de operaciones. Esto es así en cualquier análisis de datos, pero en mayor medida en el caso de la producción estadística, en la que podemos afirmar que la mayor parte del trabajo consiste precisamente en este tipo de operaciones.

Siendo esto así, existen distintas herramientas para abordar el tratamiento de datos. Una de ellas, que en nuestra organización está cada vez más presente, son precisamente un conjunto de librerías de R/tidyverse que generan scripts sencillos, permiten industrializar e integrar los procesos, así como obtener una respuesta eficiente en tiempos de procesamiento.

El objetivo de este taller es facilitar algunos scripts que a nosotros nos han resultado útiles a la hora de acceder a información, y también para transformarla. Se trata de un muestrario de soluciones que nosotros hemos implementado en algún proyecto y que pensamos que, aunque sea en ámbitos temáticos distintos, o como fase previa a la modelización, a otros usuarios también les pueden resultar útiles.

¿Presentas la comunicación a premio?

Afiliación (del autor)

Instituto de Estadística y Cartografía de Andalucía (IECA)

Autor primario: PLANELLES, Joaquin (Instituto de Estadística y Cartografía de Andalucía)

Coautor: Sra. PADILLA SÁNCHEZ, Isabel (Instituto de Estadística y Cartografía de Andalucía)

Clasificación de pistas: Estadística