

Una librería para procesar datos públicos. Presentación del paquete eph

Anónimo

Palabras clave: open data - desarrollo de paquetes - encuestas a hogares

Introducción

El trabajo con datos producidos por fuentes públicas suele encontrarse con varios problemas: uno de los más comunes es la falta de continuidad en la publicación de bases de datos. En ese sentido, la Encuesta Permanente de Hogares -EPH- del Instituto Nacional de Estadística y Censos (INDEC) en Argentina constituye una excepción. En efecto, se trata de un programa que ha publicado bases de datos usuarias con información desde 1974. No obstante, esto se ha hecho de forma “poco replicable”: desde cambios en los formatos de su publicación (dbase, .txt, .xls, .sav, etc.) hasta renombrado de algunas variables y recodificaciones de sus categorías que los hacen poco prácticos para su uso y procesamiento continuo. La inexistencia de una API para la divulgación de la información producida por INDEC genera un límite a las capacidades de procesamiento de la información, reduciendo el público usuario a i) expertos temáticos con conocimiento respecto a cómo acceder a las fuentes y ii) medios de comunicación que acceden a la información ya procesada bajo la forma de comunicados. Esto limita el potencial valor del enorme trabajo elaborado en el instituto, al desincentivar el uso de usuarios con conocimientos limitados de las fuentes, pero con capacidades de procesamiento de datos, como es la comunidad de usuaries de R. A su vez, ciertos indicadores clave presentados por la EPH cuentan con anexos metodológicos, pero no con implementaciones públicas que le permitan al público hacer uso de la metodología por fuera de los informes elaborados por el instituto.

En el marco del presente proyecto se creó el repositorio [HOLATAM](#) que busca darle un marco regional. Allí puede encontrarse una [herramienta similar](#) realizada para la encuesta Encuesta de Caracterización Socioeconómica Nacional (CASEN) de Chile. A su vez, la Encuesta Continua de Hogares de Uruguay (ECH) cuenta con un paquete [ech](#) que ha replicado la estructura de alguna de las funciones del paquete eph.

Descripción general de la librería

En este contexto, la librería [eph](#) tiene como objetivo facilitar el trabajo de aquellas usuaries de la Encuesta Permanente de Hogares - INDEC que deseen procesar datos de la misma mediante el lenguaje de programación R. La librería presenta las siguientes funcionalidades:

- una sintaxis unificada para la descarga, etiquetado y construcción de datasets con información cross-sectional comparables
- una implementación del cálculo de indicadores (pobreza) utilizando la metodología oficial

Algunas de sus funciones son:

- `get_microdata()`: Descarga las bases de microdatos
- `organize_panels()`: Permite armar un pool de datos en panel de la EPH continua,
- `organize_cno()`: Clasifica las ocupaciones según el CNO 2001
- `'organize_caes()'`: Clasifica las actividades económicas según CAES Mercosur 1.0 y CAES Mercosur
- `organize_labels()`: Etiqueta las bases siguiendo el último diseño de registro
- `map_agglomerates()`: Mapa de indicadores por aglomerado

El paquete también cuenta con otros sets de datos que pueden ser útiles para el trabajo con la EPH: algunos diccionarios que contienen la codificación de las variables geográfica s(como regiones o aglomerados) o la posición geográfica (centroides) de los aglomerados en que se releva la encuesta.

Formato de la presentación

- Presentación de los objetivos de la librería (2 min)
- Mención de algunas de sus funciones más importantes (3 min)
- Desarrollo de un workflow de ejemplo (10 min)