{ARcenso}: primeros pasos desarrollando un paquete en comunidad con rOpenSci

Andrea Gomez Vargas & Emanuel Ciardullo

Palabras clave: R, paquete, censo, Estadísticas de población, Argentina, comunidad.

En el marco del programa de Campeones y Campeonas de rOpenSci[1] y con el objetivo de identificar, reconocer y recompensar a los miembros apasionados de la comunidad que ayudan a la comunidad a crecer y mejorar, nace el proyecto de crear el paquete {ARcenso}.

{ARcenso}[2] Es un paquete en desarrollo que permitirá acceder a los datos oficiales de los censos nacionales de población en Argentina provenientes del Instituto Nacional de Estadísticas y Censos - INDEC.

Actualmente los resultados históricos censales[3] de 1970, 1980, 1991, 2001, 2010 y 2022 están disponibles en distintos formatos a través de libros físicos, PDFs, archivos en formato excel o en REDATAM[4], sin contar con un sistema o formato unificado que permita trabajar con los datos de estos seis periodos censales como base de datos (Figura 1). Además, la presentación de los datos no está homogeneizada entre periodos dificultando la comparación histórica o en serie de la información disponible.

Figura 1. Cuadro censal del año 1970 disponible para descargar en la web de INDEC.

Cuadro 7. Total del país. Población de 10 y más años, por grupo y años simples de edad, según condición de alfabetismo y sexo. Año 1970

Grupo y años simples de edad	Total			Condición de alfabetismo					
				Alfabetos			Analfabetos		
	Total	Varones	Mujeres	Total	Varones	Mujeres	Total	Varones	Mujeres
Total	**18,737,750**	**9,257,000**	**9,480,750**	**17,411,350**	**8,669,750**	**8,741,600**	**1,326,400**	**587,250**	**739,150**
10-14	2,201,150	1,114,300	1,086,850	2,100,600	1,059,400	1,041,200	100,550	54,900	45,650
10	448,400	224,700	223,700	417,300	208,250	209,050	31,100	16,450	14,650
11	444,800	229,000	215,800	421,700	215,300	206,400	23,100	13,700	9,400
12	443,300	223,000	220,300	424,600	213,050	211,550	18,700	9,950	8,750
13	433,900	220,250	213,650	419,300	212,200	207,100	14,600	8,050	6,550
14	430,750	217,350	213,400	417,700	210,600	207,100	13,050	6,750	6,300
15-19	2,098,700	1,058,850	1,039,850	2,012,900	1,013,800	999,100	85,800	45,050	40,750
15	431,150	216,100	215,050	416,750	209,150	207,600	14,400	6,950	7,450
16	417,050	209,800	207,250	399,300	200,250	199,050	17,750	9,550	8,200

[1] Introducing rOpenSci Champions - Cohort 2023-2024
[2] https://github.com/SoyAndrea/arcenso
[3] Censo Nacional de Población, Hogares y Viviendas
[4] REDATAM es un software desarrollado por la CEPAL (Comisión Económica para América Latina Y el Caribe) y utilizado ampliamente en los países de América para la difusión de estadísticas censales.

En este sentido contar con un paquete de información censal permitirá al sector público y privado, a los ciudadanos y a otros actores de la sociedad acceder a información actual e histórica sobre la población, los hogares y las viviendas de Argentina de una manera más accesible amplificando el uso de software libre.

El proyecto {ARcenso} consiste en generar un paquete que sirva de repositorio de los datos censales de estos seis periodos, ya homogeneizados en un formato de datos ordenados (tidy data) listos para su uso en R. Disponer de un paquete con estas características requiere de un armado metodológico y de diseño conceptual previo para la clasificación de los cuadros censales, lo que implica un extenso proceso 1:1 de revisión, consolidación y conversión de cuadros en excel a tabulados en R por cada indicador, por cada referencia geográfica (total del país y cada una de las 24 jurisdicciones) y por cada año censal, solo para el año 1970 se trabajaron con 357 cuadros en formato excel.

Para obtener los cuadros en este formato de datos ordenados (tidy data) (Figura 2), el paquete busca además aportar funciones elaboradas en R base para la selección de información censal, simplificar la comparabilidad y utilización conjunta de la información de los seis años censales, y de otra función que ejecute un tablero de datos soportado por los paquetes de *shiny, quarto y gt* para la exploración y consulta amigable de la información disponible.

Figura 2. Tabulado en formato "tidy data" del cuadro censal presentado en la figura 1

c70_total_del_pais_poblacion_c7

Filter

	sexo	grupo_de_edad	alfabetismo	poblacion
1	Total	10-14	Total	2201150
2	Total	10-14	Alfabetos	2100600
3	Total	10-14	Analfabetos	100550
4	Total	15-19	Total	2098700
5	Total	15-19	Alfabetos	2012900
6	Total	15-19	Analfabetos	85800
7	Total	20-24	Total	1950500
8	Total	20-24	Alfabetos	1867400
9	Total	20-24	Analfabetos	83100
10	Total	25-29	Total	1702700
11	Total	25-29	Alfabetos	1624400
12	Total	25-29	Analfabetos	78300
13	Total	30-34	Total	1580350
14	Total	30-34	Alfabetos	1501700
15	Total	30-34	Analfabetos	78650
16	Total	35-39	Total	1546400

En esta presentación nos proponemos compartir la experiencia de armar un paquete desde cero acompañados por rOpenSci y toda su comunidad, con las capacitaciones, la mentoría 1:1 y la sesiones de coworking que delimitaron los puntos de partida y estrategias para el armado del proyecto, la toma de decisiones, el armado conceptual, los desafíos y aprendizajes a la hora de construir un paquete, las primeras funciones y la hoja de ruta para incorporar los seis periodos censales.

Paquete en desarrollo en https://github.com/SoyAndrea/arcenso