

Package ‘censo2022arg’

May 8, 2026

Type Package

Title Extraction and Analysis of 2022 Argentina Census Microdata from REDATAM Databases

Version 1.0.1

Description Provides tools to extract, label, and read microdata from the 2022 National Census of Population, Households and Dwellings of Argentina stored in REDATAM databases officially distributed by INDEC. Implements a complete province-by-province extraction pipeline with efficient memory management, reconstruction of hierarchical identifiers, automatic variable labeling from official INDEC dictionaries, and integrity verification against published totals. Allows working with census data directly in R without knowledge of REDATAM syntax, and supports export to multiple formats including Parquet, CSV, SPSS and SAS. Census data must be downloaded directly from the official INDEC portal (<https://www.indec.gob.ar>). This package does not distribute census data. Duran (2026) <[doi:10.5281/zenodo.19560728](https://doi.org/10.5281/zenodo.19560728)>.

License GPL (>= 3)

Encoding UTF-8

Language es

Depends R (>= 4.1.0)

Imports redatamx, callr (>= 3.7.0), arrow, data.table, haven, readxl, dplyr

Suggests tibble, testthat (>= 3.0.0), knitr, rmarkdown

LinkingTo cpp11

VignetteBuilder knitr

URL <https://github.com/RodriDuran/censo2022arg>

BugReports <https://github.com/RodriDuran/censo2022arg/issues>

RoxygenNote 7.3.3

NeedsCompilation yes

Author Rodrigo Javier Duran [aut, cre] (ORCID: <https://orcid.org/0000-0002-3957-9510>),
Jaime Salvador [ctb] (Autor original de red_execute.cpp, red_initialize.cpp y redengine_c.h (paquete redatamx)),
CELADE - Centro Latinoamericano y Caribeno de Demografia [cph] (Titular de derechos sobre el motor REDATAM)

Maintainer Rodrigo Javier Duran <rodrigo.duran@exa.unsa.edu.ar>

Repository CRAN

Date/Publication 2026-04-21 18:42:24 UTC

Contents

censo_configurar	2
censo_descargar	3
censo_etiquetar	4
censo_info	6
censo_leer	6
censo_verificar_engine	8
extraer_redatam	9
extraer_rxdb	11

Index 13

censo_configurar	<i>Configurar el directorio de datos del Censo 2022</i>
------------------	---

Description

Define el directorio donde se guardaran todos los archivos del censo: bases REDATAM, metadatos y microdatos extraidos. Es el primer paso antes de usar cualquier otra funcion del paquete.

El directorio puede estar en cualquier ubicacion - disco interno, externo o de red. El paquete en si se instala en la libreria de R del sistema; los datos quedan donde usted elija.

Al ejecutarse, la funcion crea automaticamente la estructura de carpetas necesaria y copia los archivos de control oficial del INDEC.

Usage

```
censo_configurar(dir, persistent = FALSE)
```

Arguments

dir	Ruta al directorio raiz donde se guardaran todos los archivos del censo. Debe ser elegida por el usuario; el paquete no asume ninguna ubicacion por defecto.
persistent	Logico. Si TRUE, guarda la configuracion en .Rprofile para que el directorio quede disponible en todas las sesiones futuras sin necesidad de volver a configurar. Default FALSE.

Value

La ruta configurada (invisible).

See Also

[censo_info](#), [censo_descargar](#), [censo_verificar_engine](#)

Examples

```
# Configurar para esta sesion unicamente
censo_configurar("/home/usuario/mis_datos/censo2022")

# Configurar y guardar para todas las sesiones futuras
censo_configurar("/home/usuario/mis_datos/censo2022", persistent = TRUE)

# En Windows
censo_configurar("D:/Datos/Censo2022", persistent = TRUE)

# En un disco externo (Mac/Linux)
censo_configurar("/Volumes/MiDisco/Censo2022", persistent = TRUE)

# Ejemplo ejecutable:
old <- getOption("censo2022.dir")
censo_configurar(tempdir())
options(censo2022.dir = old) # restaurar estado anterior
```

censo_descargar

Descargar bases y documentacion del Censo 2022 desde el INDEC

Description

Descarga automaticamente desde el portal oficial del INDEC las bases de datos REDATAM, los diccionarios de variables, los cuestionarios y la documentacion metodologica del Censo 2022.

Los archivos se guardan en la estructura de directorios configurada previamente con `censo_configurar()`.

Usage

```
censo_descargar(que = "todo", overwrite = FALSE)
```

Arguments

que Character. Que descargar. Opciones:

- "todo" (default): descarga todo lo disponible
- "bases": solo las bases REDATAM (~500 MB)
- "metadatos": diccionarios de variables y documentacion

- "cuestionarios": formularios del censo en PDF
- "metodologia": documentos metodologicos en PDF

Se pueden combinar: c("cuestionarios", "metodologia").

overwrite Logico. Si TRUE, vuelve a descargar aunque el archivo ya exista. Default FALSE.

Details

La descarga de las bases puede demorar varios minutos dependiendo de la conexion. Si el enlace no esta disponible, la funcion indica como descargar los archivos manualmente desde el sitio del INDEC.

Value

Invisible NULL. Se ejecuta por sus efectos secundarios (descarga de archivos al directorio configurado).

See Also

[censo_configurar](#), [extraer_redatam](#)

Examples

```
## Not run:
# Descargar todo (recomendado la primera vez)
censo_descargar()

# Solo los metadatos (mas rapido, util para probar la configuracion)
censo_descargar(que = "metadatos")

# Forzar re-descarga de las bases
censo_descargar(que = "bases", overwrite = TRUE)

## End(Not run)
```

censo_etiquetar

Aplicar etiquetas a los microdatos extraidos del Censo 2022

Description

Aplica las etiquetas del diccionario oficial del INDEC a los archivos de microdatos generados por `extraer_redatam()`. El proceso realiza tres transformaciones sobre cada archivo:

1. Limpia los nombres de columna eliminando los sufijos numericos del motor REDATAM (p01_0 -> P01).
2. Agrega una etiqueta descriptiva a cada variable (atributo label).
3. Convierte las variables categoricas a factor con las etiquetas de sus categorias (ej: 1 -> "Jefa o jefe").

Las variables geograficas identificadoras no se convierten a factor porque sus codigos no corresponden al nivel jerarquico del diccionario.

El proceso es idempotente: si un archivo ya fue etiquetado, se detecta automaticamente y se omite sin modificarlo.

Usage

```
censo_etiquetar(
  fuente_meta = "xls",
  provincias = "all",
  bases = "all",
  etiquetar = "todo"
)
```

Arguments

fuelle_meta	Character. Fuente de los metadatos de etiquetado. "xls" (default): usa los archivos XLS del INDEC, mas rapido y sin consumo adicional de memoria. "redatam": extrae las etiquetas directamente del motor REDATAM via subprocesso, util cuando no se dispone de los archivos XLS.
provincias	Numerico o "all". Codigos de provincia a etiquetar. Default "all".
bases	Character o "all". Bases a etiquetar: "Personas", "Hogares", "Viviendas", "colectivas", "PO_VP". Default "all".
etiquetar	Character. Que etiquetas aplicar: "todo" (default), "variables" (solo nombres), "valores" (solo categorias).

Value

Invisible NULL. Modifica los archivos en el directorio configurado.

See Also

[extraer_redatam](#), [censo_leer](#)

Examples

```
## Not run:
# Etiquetar todas las provincias (configuracion automatica)
censo_etiquetar()

# Solo Formosa y Salta
censo_etiquetar(provincias = c(34, 66))

# Solo la base de Personas
censo_etiquetar(bases = "Personas")

# Usando el motor REDATAM como fuente de metadatos
censo_etiquetar(fuente_meta = "redatam")

## End(Not run)
```

`censo_info`*Mostrar el estado actual de la configuracion*

Description

Presenta un resumen del estado del paquete: directorio configurado, archivos disponibles, provincias extraidas y proximos pasos sugeridos. Tambien verifica si el motor de extraccion esta correctamente preparado.

Es util para orientarse al retomar el trabajo despues de un tiempo o para diagnosticar problemas de configuracion.

Usage`censo_info()`**Value**

La ruta del directorio configurado (invisible), o NULL si no hay configuracion activa.

See Also

[censo_configurar](#), [censo_descargar](#)

Examples

```
if (nzchar(getOption("censo2022.dir", ""))) {
  censo_info()
}
```

`censo_leer`*Leer microdatos del Censo 2022*

Description

Lee los microdatos extraidos del Censo 2022 y los devuelve como un objeto de R listo para analizar. Permite seleccionar provincias, variables y aplicar filtros de filas sin necesidad de cargar todo el conjunto de datos en memoria.

Se recomienda ejecutar `censo_etiquetar()` antes de usar esta funcion para disponer de nombres de variable legibles y categorias correctamente etiquetadas.

Usage

```
censo_leer(
  base = "Personas",
  provincias = "all",
  columnas = NULL,
  filtro = NULL,
  formato = "data.frame",
  dir = NULL
)
```

Arguments

base	Character. Base de datos a leer: <ul style="list-style-type: none"> • "Personas" (default): una fila por persona • "Hogares": una fila por hogar • "Viviendas": una fila por vivienda particular • "colectivas": personas en viviendas colectivas • "PO_VP": base completa combinada (VP + PO)
provincias	Numerico, character o "all". Provincias a leer. Acepta codigos numericos (c(66, 38)), nombres o partes del nombre (c("salta", "jujuy")), o "all" (default) para todas las provincias disponibles. "all" (default), se solicita confirmacion dado el volumen de datos (~12-15 GB en RAM para el pais completo).
columnas	Character o NULL. Nombres de variables a leer. Si es NULL (default), se cargan todas las columnas. Ejemplo: c("P01", "EDAD", "CONDUCT").
filtro	Expresion o NULL. Condicion para filtrar filas antes de cargarlas en memoria. Se evalua sobre el dataset completo sin necesidad de cargarlo. Ejemplo: quote(EDAD >= 18).
formato	Character. Formato del objeto devuelto: "data.frame" (default), "data.table" o "tibble".
dir	Character o NULL. Directorio raiz de los datos. Si es NULL (default), usa el directorio configurado con censo_configurar().

Details**## Filtros y seleccion de columnas**

Los filtros y la seleccion de columnas se aplican usando Apache Arrow antes de cargar los datos en la memoria de R. Esto permite trabajar con subconjuntos de datos sin necesidad de disponer de la RAM suficiente para el conjunto completo.

Estimacion de uso de memoria

Los archivos en disco estan comprimidos (formato parquet). Al cargarlos en R, el uso de RAM es aproximadamente 4-5 veces mayor que el tamano del archivo. Una provincia mediana ocupa ~50-100 MB en RAM. El pais completo puede superar los 12 GB.

Value

Un data.frame, data.table o tibble segun el parametro formato.

See Also

[extraer_redatam](#), [censo_etiquetar](#), [censo_configurar](#)

Examples

```
## Not run:
# Personas de Formosa
df <- censo_leer(base = "Personas", provincias = 34)

# Por nombre de provincia
df <- censo_leer(base = "Personas", provincias = c("salta", "jujuy"))

# Personas de Salta y Jujuy, solo algunas variables
df <- censo_leer(
  base      = "Personas",
  provincias = c(66, 38),
  columnas  = c("P01", "EDAD", "CONDUCT", "IDRADIO")
)

# Personas mayores de 18 anos de Cordoba
df <- censo_leer(
  base      = "Personas",
  provincias = 14,
  filtro    = quote(EDAD >= 18)
)

# Hogares de todo el pais como data.table
dt <- censo_leer(base = "Hogares", formato = "data.table")

## End(Not run)
```

censo_verificar_engine

Verificar el estado del motor de extraccion REDATAM

Description

Verifica si el motor de extraccion esta correctamente preparado para trabajar con el Censo 2022. Si no lo esta, muestra las instrucciones paso a paso para prepararlo segun su sistema operativo.

Por que es necesario este paso?

El motor REDATAM (distribuido con el paquete `redatamx`) tiene un limite interno que restringe la extraccion a 100 registros por consulta. Este limite fue disenado para uso interactivo del software REDATAM, no para la extraccion masiva de microdatos. Para poder extraer los 44 millones de registros del censo, es necesario ampliar ese limite aplicando una modificacion puntual al archivo del motor.

Esta modificacion es de bajo nivel (unos pocos bytes en el binario compilado) y no afecta ninguna otra funcionalidad. Debe realizarse una sola vez, y puede revertirse en cualquier momento usando la copia de seguridad que la funcion indica crear.

Usage

```
censo_verificar_engine()
```

Value

Invisible TRUE si el motor esta listo, FALSE si requiere preparacion.

See Also

[censo_configurar](#), [extraer_redatam](#)

Examples

```
censo_verificar_engine()
```

extraer_redatam	<i>Extraer microdatos del Censo Nacional de Poblacion, Hogares y Viviendas 2022</i>
-----------------	---

Description

Extrae los microdatos completos del Censo 2022 desde las bases REDATAM del INDEC, provincia por provincia, con verificacion automatica de integridad contra los totales oficiales publicados.

Para cada provincia genera cinco archivos:

- provincia_PO_VP.parquet - base completa (una fila por persona)
- provincia_Personas.parquet - variables de persona
- provincia_Hogares.parquet - variables de hogar
- provincia_Viviendas.parquet - variables de vivienda
- provincia_colectivas.parquet - viviendas colectivas

La extraccion es retomable: si se interrumpe, al volver a ejecutar continua desde donde quedo sin repetir lo ya procesado.

Usage

```
extraer_redatam(  
  provincias = "all",  
  formatos = "parquet",  
  max_por_bloque = 10,  
  dic_path_vp = NULL,  
  dic_path_po = NULL,  
  dic_path_vc = NULL,  
  output_dir = NULL,  
  control_dir = NULL  
)
```

Arguments

provincias	Numerico o "all". Codigos de provincia a extraer. Ejemplo: c(34, 66) extrae Formosa y Salta. Con "all" (default) procesa las 24 provincias.
formatos	Character. Formato de salida de los archivos. Default "parquet". Tambien acepta "csv", "sav", "sas" o combinaciones: c("parquet", "csv").
max_por_bloque	Integer. Numero maximo de variables por bloque de extraccion. Default 10. Reducir en equipos con poca memoria RAM.
dic_path_vp	Character. Ruta al archivo .rxdb de la Base VP. Si es NULL (default), se usa la ruta configurada con censo_configurar().
dic_path_po	Character. Ruta al archivo .rxdb de la Base PO. Si es NULL (default), se usa la ruta configurada.
dic_path_vc	Character. Ruta al archivo .rxdb de la Base VC. Si es NULL (default), se usa la ruta configurada.
output_dir	Character. Directorio de salida para los microdatos. Si es NULL (default), se usa el directorio configurado.
control_dir	Character. Directorio con los archivos de control del INDEC para verificacion. Si es NULL (default), se usa el directorio configurado.

Details

Bases del Censo 2022

El INDEC distribuyo tres bases complementarias:

- **VP** (Viviendas Particulares): contiene las variables principales e incluye los niveles geograficos de fraccion y radio censal, necesarios para analisis a nivel de radio.
- **PO** (Pueblos Originarios, Afrodescendientes e Identidad de Genero): agrega seis variables sensibles no incluidas en VP. No incluye fraccion ni radio censal.
- **VC** (Viviendas Colectivas): personas en establecimientos colectivos (hospitales, carceles, etc.).

La base extraida combina VP y PO, obteniendo el radio censal de VP y las seis variables adicionales de PO.

Gestion de memoria

Cada bloque se extrae en un subproceso Rscript independiente. Esto evita la acumulacion de memoria del motor REDATAM en la sesion principal de R. El proceso maestro solo lee los archivos intermedios desde disco y los une en memoria.

Value

Invisible. Una lista con los elementos: fallidas (provincias con errores) y log (ruta al log).

See Also

[censo_configurar](#) para configurar las rutas, [censo_etiquetar](#) para etiquetar los archivos generados, [censo_leer](#) para leer los microdatos extraidos.

Examples

```
## Not run:
# Extraer todas las provincias
extraer_redatam()

# Solo Formosa (prueba rapida)
extraer_redatam(provincias = 34)

# Salta y Jujuy, exportando tambien a CSV
extraer_redatam(provincias = c(66, 38), formatos = c("parquet", "csv"))

## End(Not run)
```

 extraer_rxdb

Extraer microdatos de cualquier base RedatamX (.rxdb)

Description

Convierte una base de datos en formato RedatamX (.rxdb) a un archivo de salida con todos sus microdatos, sin asumir ninguna estructura predefinida de variables, entidades ni jerarquia geografica.

La funcion detecta automaticamente la entidad mas granular de la jerarquia (entidad hoja), construye bloques de variables dinamicamente desde el diccionario, y extrae cada bloque en un sub-proceso independiente para una gestion eficiente de la memoria RAM.

Para el Censo Argentina 2022, utilice [extraer_redatam](#), que incluye filtrado provincial, reconstruccion de identificadores jerarquicos y verificacion de integridad.

Usage

```
extraer_rxdb(
  dic_path,
  output_file = NULL,
  max_por_bloque = 10,
  formato = "parquet",
  verbose = TRUE
)
```

Arguments

dic_path	Character. Ruta al archivo de diccionario (.rxdb).
output_file	Character. Ruta al archivo de salida. Si es NULL (default), se genera automaticamente en el mismo directorio que dic_path con el nombre de la base.
max_por_bloque	Integer. Numero maximo de variables por bloque de extraccion. Default 10. Reducir en equipos con poca memoria RAM.
formato	Character. Formato de salida: "parquet" (default) o "csv".
verbose	Logico. Si TRUE (default), muestra el progreso detallado en consola.

Details**## Requisito previo**

Requiere el motor RedatamX con el limite de extraccion ampliado. Verifique el estado del motor con `censo_verificar_engine()` antes de usar esta funcion.

Compatibilidad

Compatible con bases en formato **.rxdb** (RedatamX). No es compatible con el formato Redatam7 (.dicx + .rbf), utilizado por censos como el de Ecuador 2022.

Gestion de memoria

La extraccion se realiza en bloques de `max_por_bloque` variables. Cada bloque se procesa en un subproceso Rscript independiente que libera la memoria del motor C++ al finalizar. El proceso principal solo une los bloques leidos desde disco.

Países con bases RedatamX conocidas (ronda 2020)

- **Argentina 2022:** usar `extraer_redatam()` para el pipeline completo.
- **Guatemala, Mexico, Bolivia:** compatibilidad pendiente de verificacion.

Value

Invisible. La ruta al archivo generado.

See Also

[extraer_redatam](#) para el Censo Argentina 2022, [censo_verificar_engine](#) para verificar el motor.

Examples

```
## Not run:
# Uso basico: convertir cualquier base .rxdb a parquet
extraer_rxdb(
  dic_path    = "/ruta/a/base.rxdb",
  output_file = "/ruta/salida/microdatos.parquet"
)

# Con bloques mas pequenos para equipos con poca memoria RAM
extraer_rxdb(
  dic_path      = "/ruta/a/base.rxdb",
  output_file   = "/ruta/salida/microdatos.parquet",
  max_por_bloque = 5
)

## End(Not run)
```

Index

`censo_configurar`, [2](#), [4](#), [6](#), [8–10](#)
`censo_descargar`, [3](#), [3](#), [6](#)
`censo_etiquetar`, [4](#), [8](#), [10](#)
`censo_info`, [3](#), [6](#)
`censo_leer`, [5](#), [6](#), [10](#)
`censo_verificar_engine`, [3](#), [8](#), [12](#)

`extraer_redatam`, [4](#), [5](#), [8](#), [9](#), [9](#), [11](#), [12](#)
`extraer_rxdb`, [11](#)