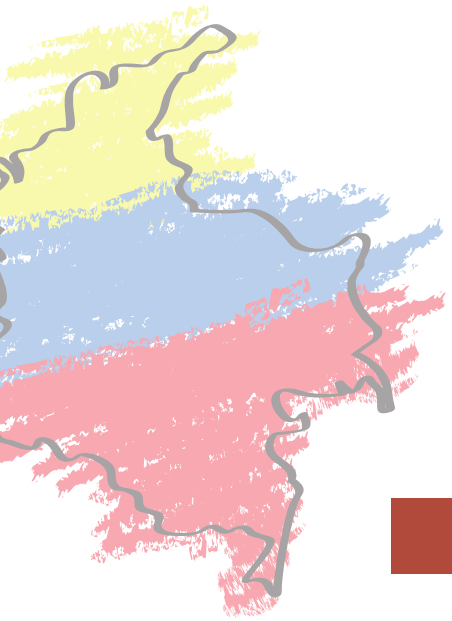

PULSO SOCIAL COLOMBIA

REALIDADES Y PERSPECTIVAS TERRITORIALES

7 de noviembre de 2021

Protocolo de datos

BID
Mejorando vidas

UNIVERSIDAD EAFIT

Autores

Equipo BID:	Priscilla Grissel Gutierrez Juarez
Equipo Universidad EAFIT	Mónica Hernandez y Juan Carlos Muñoz
Investigadores Junior:	Germán Angulo y Ana Pirela
Versión:	1.0

1 Protocolo de datos Pulso Social

Este documento detalla el protocolo de manejo de datos siguiendo la política de datos abiertos y buenas prácticas del Banco Interamericano de Desarrollo (BID, 2018), el flujo de trabajo de Humans Rights Data Analysis Group (HRDAG) y las buenas prácticas para el manejo de códigos y datos en las ciencias sociales de Gentzkow & Shapiro (2014).

1.1 Datos y códigos abiertos

Para garantizar la integridad, accesibilidad y replicabilidad de los datos empleados en el análisis y los resultados obtenidos, todos los datos y códigos se encuentran disponibles en el repositorio de GitHub “BID_Pulso_Social”. Los datos utilizados son escogidos y procesados de forma que se garantiza que sean exhaustivos e interoperables:

- Los datos son exhaustivos: se emplean bases de datos de fuentes oficiales desagregadas al menor nivel posible (género, grupo etario, raza/etnia, territorial).
- Los datos son interoperables: las bases de datos son procesadas y guardadas en una estructura estándar que permite combinar distintos tipos de datos.

Adicionalmente, las bases de datos con las que se elaboran los resultados y análisis finales son guardadas en formato .csv, de manera que sean reconocibles por múltiples herramientas en varios lenguajes de programación.

1.2 Flujo de trabajo

1.2.1 Directorios generales

El análisis empírico consiste en el procesamiento de datos y construcción de estadísticas descriptivas, organizados en dos directorios: "Dataz "Descriptives".

- "Dataçontiene el procesamiento de las bases de datos originales para construir bases estandarizadas que servirán de insumo para las estadísticas descriptivas ("Descriptives").
- "Descriptivesçontiene códigos que usan las bases estandarizadas de "Data"para elaborar bases, gráficas, tablas y demás elementos para el análisis descriptivo.

1.2.2 Procesamiento de datos

El procesamiento de una fuente de datos concreta es una tarea. Cada tarea consta de 4 carpetas: Docs, Input, Output y Src (source). Las carpetas funcionan de la siguiente forma:

- La documentación de los datos, diccionarios y otros tipos de documentos de referencia están en "Docs/"
- Los datos originales por leer estan en Ïnput/"
- Los códigos que procesan los datos originales, y construyen bases, graficas, tablas y demás, están en "Src/"
- Los resultados del procesamiento (bases, gráficas, tablas) están en .ºutput/"

Los archivos de Ïnput/"son únicamente de lectura, nunca deben sobrescribirse. En lo posible, tampoco deben cambiarse sus nombres originales o formato, de manera que el código los lea tal cual como fueron descargados de la página web o fuente, sin pasos intermedios. Esto permite replicar los resultados con mayor facilidad. Los códigos de "Src/"leen los archivos en Ïnput/"para crear bases de datos, gráficas, tablas y demás elementos que serán guardados en .ºutput/". Los códigos también pueden llamar elementos de .ºutput/": por ejemplo, tras guardar una base de datos, otro código puede llamarla para construir gráficas. Los resultados de los códigos de "Src/"son guardados en .ºutput/". La Figura 1 muestra un ejemplo del flujo de trabajo para una base de datos en particular, “Ejemplo_base_datos”. El procesamiento de esta base de datos se encuentra en el directorio “Data/Ejemplo_base_datos/” y se realiza con los códigos de “Src/” a partir de la lectura de los datos originales de “Input/”, y los resultados se guardan en “Output/”. Figura 1. Flujo de trabajo

1.2.3 Tareas autocontenidas y autodocumentadas

Al organizar cada tarea en las 4 carpetas, la tarea se convierte en autocontenida y autodocumentada:

- La tarea está autocontenida porque todas las transformaciones relacionadas con esa base de datos, desde el procesamiento del dato original (raw), hasta la construcción de otras bases, tablas y gráficas, se desarrolla y guarda en un mismo directorio: "Data/Ejemplo_base_datos/
- La tarea está autodocumentada porque todos los documentos y diccionarios relacionados con ella están en la carpeta "Docs/ todos los códigos están en "Src/". Todas las transformaciones realizadas a los datos puros están en "Src/ no se realizan cambios en hojas de cálculo de excel ni otros tipos de software que dificulten la trazabilidad en la construcción de los resultados finales.

1.2.4 Construcción de estadísticas descriptivas

El directorio “Descriptives” se encarga de leer las bases estandarizadas de “Data” y construir con ellas estadísticas descriptivas, como tablas, gráficas y demás elementos.

- En este directorio no se realiza procesamiento de datos puros, sino que se parte de los datos estandarizados, que pueden seguir transformándose para construir resultados.
- "Descriptives" también se compone de tareas, en las que se parte de una base de datos en particular y con ella se construyen estadísticas descriptivas. Por ejemplo, "Descriptives/Ejemplo_tasa_desempleo" tiene las carpetas "Input", "Output", y "Src".
- En "Input" se guardan bases de datos transformadas para producir gráficas, tablas y otros elementos.
- En "Output" se guardan resultados: gráficas, tablas y otros elementos de estadística descriptiva.
- "Src" contiene los códigos que crean bases de datos para las descriptivas, y que construyen gráficas, tablas y demás elementos

1.3 Manejo de códigos y datos

El manejo de datos tiene los objetivos de transformar los datos puros a un formato estándar, remover datos atípicos, corregir errores en las variables, generalizar los nombres de las bases de datos y de las variables y emparejar bases.

1.3.1 Nombres de carpetas y elementos

- Las carpetas se nombran con mayúscula en la primera letra y los espacios entre palabras se reemplazan con barra al piso. Por ejemplo: "Ejemplo_base_datos/".
- Las bases de datos, gráficas, tablas y otros resultados que provengan del código se nombran en minúsculas y los espacios entre palabras se reemplazan con barra al piso. Por ejemplo: "base_tasa_desempleo_municipal.rds", y "grafica_tasa_desempleo_municipal.jpeg".
- Las bases de datos, gráficas, tablas y otros que quieran hacer énfasis en los periodos separan los años con un guion. Por ejemplo: "base_tasa_desempleo_municipal_2010-2012.rds", "base_tasa_desempleo_municipal_2012-2020.rds", "grafica_tasa_desempleo_municipal_201 2020.jpeg".
- Las bases de datos, gráficas, tablas y demás elementos construidos con el código indican qué tipo de elemento es (una base de datos, una gráfica, una tabla), una descripción de su contenido (desempleo, pobreza, educación), el nivel de desagregación (nacional, departamental, municipal), y cuando es necesario, el periodo para el que están disponibles los datos (2010-2020, 2012-2014).

1.3.2 Nombres de códigos y sintaxis

- Los códigos se nombran siguiendo un orden alfanumérico que permite identificar el orden en que se ejecutan. Por ejemplo: 01a_procesar_geih.R, 01b_procesar_ecv.R, 02a_indicadores_geih.R, 02b_indicadores_ecv.R.
- Cada código incluye una descripción corta de las acciones que realiza y la fecha en que fue editado por última vez.
- Las variables y objetos de los códigos se nombran en minúsculas con barra al piso para separar palabras. Por ejemplo: data_desempleo, tasa_desempleo.