05. - Data Collection_CU_04_01_zonasgeo_v_01

June 8, 2023

#

CU04_Optimización de vacunas

Citizenlab Data Science Methodology > II - Data Processing Domain *** > # 05.- Data Collection

Data Collection is the process to obtain and generate (if required) necessary data to model the problem.

0.0.1 01. Zonas básicas de salud de shp a json

Table of Contents

Settings

Data Load

ETL Processes

Import data from: CSV, Excel, Tab, JSON, SQL, and Parquet files

Synthetic Data Generation

Fake Data Generation

Open Data

Data Save

Main Conclusions

Main Actions

Acciones done

Acctions to perform

0.1 Settings

0.1.1 Encoding

Con la siguiente expresión se evitan problemas con el encoding al ejecutar el notebook. Es posible que deba ser eliminada o adaptada a la máquina en la que se ejecute el código.

[ ]:
```
Sys.setlocale(category = "LC_ALL", locale = "es_ES.UTF-8")
```

'es_ES.UTF-8/es_ES.UTF-8/es_ES.UTF-8/C/es_ES.UTF-8/C'

0.1.2 Packages to use

- {tcltk} for selecting files and paths (if needed)
- {sf} for reading and writing files with spatial information
- {dplyr} for data exploration

[15]:
```
library(tcltk)
library(sf)
library(dplyr)
```

0.1.3 Paths

[16]:
```
iPath <- "Data/Input/"
oPath <- "Data/Output/"
```

0.2 Data Load

If there are more than one input file, make as many sections as files to import.

OPCION A: Seleccionar fichero en ventana para mayor comodidad

Data load using the {tcltk} package (uncomment for using this option)

[17]:
```
# file_data_01 <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

Instrucciones - Los ficheros de entrada del proceso están siempre en Data/Input/.
- Si hay más de un fichero de entrada, se crean tantos objetos iFile_xx y file_data_xx como ficheros de entrada (xx número correlativo con dos dígitos, rellenar con ceros a la izquierda) - Comentar si se usa la opción A

[18]:
```
iFile <- "Zonas Básicas de Salud del Área Única/200001380.shp"
file_data <- paste0(iPath, iFile)

if(file.exists(file_data)){
    cat("Se leerán datos del archivo: ", file_data)
} else{
    warning("Cuidado: el archivo no existe.")
}
```

```
Se leerán datos del archivo:  Data/Input/Zonas Básicas de Salud del Área
Única/200001380.shp
```

Data file to dataframe Utilizamos en todos los notebooks SIEMPRE data como nombre de la matriz (dataframe) principal donde cargamos los datos. Si se importan más ficheros, se le pone un sufijo con número correlativo en dos cifras y una palabra informativa. Por ejemplo, data_01_poblacion. Se repiten las celdas de estructura y muestra de datos para dataframe.

[19]:
```
data <- st_read(file_data,
                options = "ENCODING=WINDOWS-1252")
```

```
options:        ENCODING=WINDOWS-1252
Reading layer `200001380' from data source
  `/Users/emilio.lcano/academico/gh_repos/__transferencia/citizenlab/notebooks/I
I_data_processing/04_vacunas/Data/Input/Zonas Básicas de Salud del Área
Única/200001380.shp'
  using driver `ESRI Shapefile'
Simple feature collection with 286 features and 3 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: 365717.6 ymin: 4415488 xmax: 495593.9 ymax: 4557518
Projected CRS: ED50 / UTM zone 30N
```

NOTE: This object is a special type of dataframe that includes the geometry of each observation.

Estructura de los datos:

[20]:
```
glimpse(data)
```

```
Rows: 286
Columns: 4
$ CODBDT    <int> 686213, 686214, 686215, 686216, 686217,
686218, 686219, 6862…
$ GEOCODIGO <chr> "001", "002", "003", "004", "005", "006",
"007", "008", "009…
$ DESBDT    <chr> "Abrantes", "Acacias", "Adelfas",
"Alameda", "Alameda de Osu…
$ geometry  <MULTIPOLYGON [m]> MULTIPOLYGON (((439140.4
44…, MULTIPOLYGON ((…
```

Muestra de datos:

[21]:
```
slice_head(data, n = 5)
```

A sf: 5 × 4

CODBDT <int>	GEOCODIGO <chr>	DESBDT <chr>	geometry <MULTIPOLYGON [m]>
686213	001	Abrantes	MULTIPOLYGON (((439140.4 44...
686214	002	Acacias	MULTIPOLYGON (((440034.4 44...
686215	003	Adelfas	MULTIPOLYGON (((443565.2 44...
686216	004	Alameda	MULTIPOLYGON (((440763.1 44...
686217	005	Alameda de Osuna	MULTIPOLYGON (((452527.5 44...

0.3 ETL Processes

0.3.1 Transform data

- Para asegurar la trazabilidad, se guardan las transformaciones en un objeto diferente con el prefijo t
- Convertir a crs 4326 que es el estándar del caso de uso (proyecto?)

[22]:
```
tdata <- data |>
  st_transform(4326)
```

Muestra datos transformados:

[23]:
```
tdata |> slice_head(n = 5)
```

A sf: 5 × 4

	CODBDT <int>	GEOCODIGO <chr>	DESBDT <chr>	geometry <MULTIPOLYGON [°]>
	686213	001	Abrantes	MULTIPOLYGON (((-3.718306 4...
	686214	002	Acacias	MULTIPOLYGON (((-3.707966 4...
	686215	003	Adelfas	MULTIPOLYGON (((-3.666363 4...
	686216	004	Alameda	MULTIPOLYGON (((-3.69947 40...
	686217	005	Alameda de Osuna	MULTIPOLYGON (((-3.561629 4...

Estructura de la projección:

[24]:
```
st_crs(tdata)
```

```
Coordinate Reference System:
  User input: EPSG:4326
  wkt:
GEOGCRS["WGS 84",
    ENSEMBLE["World Geodetic System 1984 ensemble",
        MEMBER["World Geodetic System 1984 (Transit)"],
        MEMBER["World Geodetic System 1984 (G730)"],
        MEMBER["World Geodetic System 1984 (G873)"],
        MEMBER["World Geodetic System 1984 (G1150)"],
        MEMBER["World Geodetic System 1984 (G1674)"],
        MEMBER["World Geodetic System 1984 (G1762)"],
        MEMBER["World Geodetic System 1984 (G2139)"],
        ELLIPSOID["WGS 84",6378137,298.257223563,
            LENGTHUNIT["metre",1]],
        ENSEMBLEACCURACY[2.0]],
    PRIMEM["Greenwich",0,
        ANGLEUNIT["degree",0.0174532925199433]],
    CS[ellipsoidal,2],
        AXIS["geodetic latitude (Lat)",north,
            ORDER[1],
            ANGLEUNIT["degree",0.0174532925199433]],
        AXIS["geodetic longitude (Lon)",east,
            ORDER[2],
            ANGLEUNIT["degree",0.0174532925199433]],
    USAGE[
        SCOPE["Horizontal component of 3D system."],
        AREA["World."],
        BBOX[-90,-180,90,180]],
    ID["EPSG",4326]]
```

0.4 Synthetic Data Generation

Estos datos no requieren tareas de este tipo.

0.5 Fake Data Generation

Estos datos no requieren tareas de este tipo.

0.6 Open Data

Los datos de origen fueron descargados de una fuente abierta en fichero zip que se ha descomprimido dentro de la carpeta Data/Input.

0.7 Data Save

Este proceso, puede copiarse y repetirse en aquellas partes del notebbok que necesiten guardar datos. Recuerde cambiar la extensión añadida del fichero para diferenciarlas

Identificamos los datos a guardar

[25]:
```
data_to_save <- tdata
```

Estructura de nombre de archivos:

- Código del caso de uso, por ejemplo “CU_04”
- Número del proceso que lo genera, por ejemplo ”_05”.
- Número de la tarea que lo genera, por ejemplo ”_01”
- En caso de generarse varios ficheros en la misma tarea, llevarán _01 _02 … después
- Nombre: identificativo de “properData”, por ejemplo ”_zonasgeo”
- Extensión del archivo

Ejemplo: ”CU_04_05_01_01_zonasgeo.json, primer fichero que se genera en la tarea 01 del proceso 05 (Data Collection) para el caso de uso 04 (vacunas)

Importante mantener los guiones bajos antes de proceso, tarea, archivo y nombre

0.7.1 Proceso 05

[26]:
```
caso <- "CU_04"
proceso <- '_05'
tarea <- "_01"
archivo <- ""
proper <- "_zonasgeo"
extension <- ".json"
```

OPCION A: Uso del paquete “tcltk” para mayor comodidad

- Buscar carpeta, escribir nombre de archivo SIN extensión (se especifica en el código)
- Especificar sufijo2 si es necesario
- Cambiar datos por datos_xx si es necesario
- Descomentar líneas si se usa esta opción

[27]:
```
# file_save <- paste0(caso, proceso, tarea, archivo, tcltk::tkgetSaveFile(),␣
 ↪extension)
# path_out <- paste0(oPath, file_save)
# st_write(obj = data_to_save,
        #  dsn = path_out,
        #  driver = "GeoJSON",
        #  delete_dsn = TRUE)

# cat('File saved as: ')
# path_out
```

OPCION B: Especificar el nombre de archivo

- Los ficheros de salida del proceso van siempre a Data/Output/.

[28]:
```
file_save <- paste0(caso, proceso, tarea, archivo, proper, extension)
path_out <- paste0(oPath, file_save)
st_write(obj = data_to_save,
         dsn = path_out,
         driver = "GeoJSON",
         delete_dsn = TRUE)


cat('\nFile saved as: ', path_out)
```

```
Deleting source `Data/Output/CU_04_05_01_zonasgeo.json' failed
Writing layer `CU_04_05_01_zonasgeo' to data source
  `Data/Output/CU_04_05_01_zonasgeo.json' using driver `GeoJSON'
Writing 286 features with 3 fields and geometry type Multi Polygon.

File saved as:  Data/Output/CU_04_05_01_zonasgeo.json
```

Copia del fichero a Input Si fichero es usado en otras tareas, copiar a carpeta Input

[29]:
```
path_in <- paste0(iPath, file_save)
file.copy(path_out, path_in)
```

TRUE

0.8 Main Conclusions

List and describe the general conclusions of the analysis carried out.

0.8.1 Prerequisitos

This working code needs the following conditions:

- For using the interactive selection of file, the {tcltk} package must be installed. It is not needed in production.
- The {dplyr} package must be installed. It is part of the *tidyverse*.

- The {sf} package must be installed. It needs some system requierements, check the package documentation at https://r-spatial.github.io/sf/
- The data paths `Data/Input` and `Data/Output` must exist (relative to the notebook path)

0.8.2 Gestión de la configuración

This notebook has been tested with the following versions of R and packages. It cannot be assured that later versions work in the same way: * R 4.2.2 * tcltk 4.2.2 * sf 1.0.9 * dplyr 1.0.10

0.8.3 Estructuras de datos

- Los datos geográficos en el fichero JSON vienen con la proyección CRS WGS 84 (código 4326)
- Hay 286 zonas básicas de salud de las que se dispone de:
  - GEOCODIGO, string de tamaño tres con el código de la zona básica sanitaria que después se puede unir a otras tablas.
  - DESBDT, string de tamaño variable que contiene el nombre de la zona básica sanitaria y se usará para describir la zona.
  - CODBDT, entero que no utilizamos y no está claro para qué puede servir

0.8.4 Consideraciones para despliegue en piloto

- Los datos de origen fueron obtenidos de https://gestiona.comunidad.madrid/nomecalles/DescargaBDTCorte. en fichero zip con formato .shp y descomprimido en la carpeta Data/Input
- Si la información geográfica cambia, se debería actualizar el fichero y volver a ejecutar todos los procesos.

0.8.5 Consideraciones para despliegue en producción

- Se deben crear los procesos ETL en producción necesarios para que los datos de entrada estén actualizados

0.9 Main Actions

Acciones done Indicate the actions that have been carried out in this process

- Se ha usado el encoding de Windows para la importación
- Se ha convertido la proyección a EPSG:4326 (WGS84)
- Se ha exportado a formato GeoJson

Acctions to perform Indicate the actions that must be carried out in subsequent processes

- Se deben extraer los datos sin polígonos y guardar como .csv

0.10 CODE TO DEPLOY (PILOT)

A continuación se incluirá el código que deba ser llevado a despliegue para producción, dado que se entiende efectúa operaciones necesarias sobre los datos en la ejecución del prototipo

Description

- No hay código que desplegar en el piloto, ya que estos datos son estáticos o en todo caso cambian con muy poca frecuencia, altamente improbable durante el proyecto.

- El fichero deberá ser copiado a la ruta de datos del caso de uso para ser utilizado en la representación de mapas

CODE

[30]:
```
# incluir código
```