Proyecto Visualizacion de Datos

Benito Pastor & Pedro Gil & Alejandro Hernandez & Rosa Martinez & Jorge Albalat

2023-04-14

Librerías

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
library(dplyr)
```

En este proyecto de visualización de datos vamos a analizar un conjunto de datos relacionados con la contaminación átmosferica que hay en diferentes puntos de la ciudad de Valencia.

Procedemos a realizar un análisis univariante de los datos, pero antes de ello, realizaremos un acondicionamiento de los datos, eliminando variables que no aportan información y/o que son irrelevantes para el estudio.

Para ello, leemos los datos proporcionados por el Ayuntamiento de Valencia:

1. Lectura de los datos.

2. Acondicionamiento de los datos.

Eliminamos la columna "Id" que no aporta ninguna informacion relevante, como tampoco lo hace la columna "Fecha baja" y "Fecha Creacion". Además, al realizar un pequeño análisis visual de los datos, nos damos cuenta de que las variables donde se ve el ng/m3 de cada sustancia, no contiene información para la gran mayoría de registros y por tanto procedemos a eliminarlas.

```
df <- df %>%
  select(!(c("Id", "Fecha baja","Fecha creacion"))) %>%
  select(1:26)
```

Ahora cambiamos el tipo de dato de la variable "Dia de la semana", "Dia del mes" y "Estacion" a tipo factor para poder trabajar con estas variables.

```
df <- df %>%
  mutate("Dia de la semana" = factor(df$`Dia de la semana`)) %>%
  mutate(Estacion = factor(df$Estacion)) %>%
  mutate("Dia del mes" = factor(df$`Dia del mes`))
```

Ahora eliminamos todos los registros que tengan en la variable Estacion el nombre “Nazaret Metero”, “Conselleria Meteo” y “Puerto Valencia” ya que no contienen información útil para el estudio o son registros prácticamente duplicados.

```
df <- df %>%
  filter(Estacion != "Nazaret Meteo") %>%
  filter(Estacion != "Conselleria Meteo") %>%
  filter(Estacion != "Puerto Valencia")
```

Vamos a cambiar algunos nombres de las categorías de la variable “Estacion” para que sea más fácil trabajar con ellos posteriormente.

```
df <- df %>% mutate(Estacion = recode(Estacion,
                                      "Avda. Francia" = "Francia",
                                     "Bulevard Sud" = "Boulevar Sur",
                                     "Moli del Sol" = "Molí del Sol",
                                     "Pista Silla" = "Pista de Silla",
                                     "Politecnico" = "Universidad Politécnica",
                                     "Puerto llit antic Turia" = "Puerto llit antic Turia(s)",
                                     "Puerto Moll Trans. Ponent" = "Puerto Moll Trans. Ponent(s)",
                                     "Valencia Centro" = "Centro",
                                      "Valencia Olivereta"= "Olivereta",
                                      "Viveros"= "Viveros"))
```

3. Detección de NA

```
#Calculamos el número de NA que existen en nuestro conjunto de observaciones.
num_nona <- sum(complete.cases(df) == TRUE)
num_na <- nrow(df) - num_nona
num_na
```

```
## [1] 37696
```

```
#Repetimos el proceso unicamente teniendo en cuenta las variables numéricas.
df_var_num <- df %>% select(5:26)
num_nona <- sum(complete.cases(df_var_num) == TRUE)
num_na_num <- nrow(df_var_num) - num_nona
num_na_num
```

```
## [1] 37696
```

Este resultado nos indica que todos los registros tienen al menos una variable con NA porque existe el mismo número de registros que registros cuyo número de variables con NA es al menos una y ya que estos NA se producen sobre las variables númericas.

Ya hemos acondicionado los datos, ahora realizaremos un análisis univariante:

4. Análisis Univariante.

Procedemos a realizar un pequeño análisis univariante, mediante el comando summary()

```
summary(df)
```

```
##      Fecha              Dia de la semana  Dia del mes
##  Min.   :2004-01-01   Domingo  :5381    16     : 1239
##  1st Qu.:2011-08-20   Jueves   :5387    17     : 1239
##  Median :2015-12-08   Lunes    :5381    18     : 1239
##  Mean   :2015-07-25   Martes   :5383    19     : 1239
##  3rd Qu.:2020-01-11   Miercoles:5383    20     : 1239
##  Max.   :2022-12-31   Sabado   :5391    21     : 1239
##                       Viernes  :5390    (Other):30262
##                     Estacion         PM1            PM2.5
##  Pista de Silla         :6940   Min.   : 0.00   Min.   : 0.00
##  Viveros                :6940   1st Qu.: 4.00   1st Qu.: 7.00
##  Universidad Politécnica:5479   Median : 7.00   Median :11.00
##  Francia                :5113   Mean   : 9.11   Mean   :12.49
##  Molí del Sol           :5113   3rd Qu.:12.00   3rd Qu.:16.00
##  Boulevar Sur           :4748   Max.   :71.00   Max.   :88.00
##  (Other)                :3363   NA's   :26771   NA's   :14386
##       PM10              NO              NO2             NOx
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.0   Min.   :  0.00
##  1st Qu.: 12.00   1st Qu.:  3.00   1st Qu.: 15.0   1st Qu.: 21.00
##  Median : 19.00   Median :  7.00   Median : 25.0   Median : 36.00
##  Mean   : 20.95   Mean   : 13.71   Mean   : 28.2   Mean   : 48.94
##  3rd Qu.: 27.00   3rd Qu.: 17.00   3rd Qu.: 38.0   3rd Qu.: 63.00
##  Max.   :209.00   Max.   :246.00   Max.   :129.0   Max.   :455.00
##  NA's   :9569     NA's   :4273     NA's   :4275    NA's   :4282
##        O3               SO2               CO              NH3
##  Min.   :  2.00   Min.   : 0.000   Min.   :0.000   Min.   : 0.00
##  1st Qu.: 34.00   1st Qu.: 3.000   1st Qu.:0.100   1st Qu.: 4.00
##  Median : 50.00   Median : 3.000   Median :0.200   Median : 5.00
##  Mean   : 48.55   Mean   : 3.479   Mean   :0.222   Mean   : 5.46
##  3rd Qu.: 63.00   3rd Qu.: 4.000   3rd Qu.:0.300   3rd Qu.: 7.00
##  Max.   :122.00   Max.   :24.000   Max.   :1.600   Max.   :22.00
##  NA's   :4761     NA's   :5295     NA's   :18096   NA's   :35115
##       C7H8             C6H6             Ruido            C8H10
##  Min.   : 0.00   Min.   : 0.00   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.60   1st Qu.: 0.60   1st Qu.: 58.00   1st Qu.: 0.80
##  Median : 3.30   Median : 1.20   Median : 61.00   Median : 1.70
##  Mean   : 4.53   Mean   : 1.43   Mean   : 60.06   Mean   : 2.85
##  3rd Qu.: 5.80   3rd Qu.: 1.90   3rd Qu.: 63.00   3rd Qu.: 3.70
##  Max.   :38.50   Max.   :14.60   Max.   :250.00   Max.   :33.70
##  NA's   :33708   NA's   :33440   NA's   :26888    NA's   :33416
##  Velocidad del viento Direccion del viento  Temperatura    Humidad relativa
##  Min.   : 0.100       Min.   :  0.0        Min.   : 3.10   Min.   : 14.0
##  1st Qu.: 0.800       1st Qu.: 67.0        1st Qu.:13.90   1st Qu.: 56.0
##  Median : 1.300       Median :200.0        Median :18.50   Median : 67.0
##  Mean   : 1.617       Mean   :177.2        Mean   :18.89   Mean   : 65.6
##  3rd Qu.: 2.000       3rd Qu.:270.0        3rd Qu.:24.10   3rd Qu.: 75.0
##  Max.   :13.400       Max.   :360.0        Max.   :34.20   Max.   :100.0
##  NA's   :22999        NA's   :22965        NA's   :27392   NA's   :27552
##     Presion      Radiacion solar Precipitacion     Velocidad maxima del viento
##  Min.   : 981    Min.   :  0.0   Min.   :  0.000   Min.   : 0.110
##  1st Qu.:1003    1st Qu.:119.0   1st Qu.:  0.000   1st Qu.: 3.100
```

```
##  Median :1010    Median :189.0   Median :  0.000   Median : 4.100
##  Mean   :1010    Mean   :189.6   Mean   :  1.358   Mean   : 5.247
##  3rd Qu.:1016    3rd Qu.:261.0   3rd Qu.:  0.000   3rd Qu.: 6.000
##  Max.   :1042    Max.   :429.0   Max.   :210.400   Max.   :25.600
##  NA's   :27313   NA's   :27356   NA's   :31482     NA's   :31191
```

5. Arreglo para la exportacion.

Ahora que hemos realizado este proceso, se realiza un duplicado de los datos para proceder a exportarlos y realizar una serie de modificaciones.

Dado que necesitamos unificar los registros según la variable "Estacion" para poder trabajar con QGIS y poder relacionar los datos con la capa de datos que almacena los nombres de las estaciones, para ello calculamos la media de cada una de las categorías de la variable "Estacion" y así obtener un nuevo conjunto de datos con 10 registros que se corresponderían con el número de distintas categorías que tiene la variable "Estacion".

```
df_num <- df %>%
  group_by(Estacion) %>%
  summarize(across(where(is.numeric), ~ mean(., na.rm = TRUE)))
```

Analizando los nombres de las de las distintas estaciones del data frame de R y el conjunto de datos de QGIS, nos damos cuenta de que hay 2 nombres de estaciones en R que no se encuentran en QGIS y que hay 3 nombres en QGIS que no se encuentran en R, por tanto hay únicamente 8 que son comunes.

A continuación un resumen de lo dicho:

- Número de nombres que se encuentran en R y no en QGIS: 2 (CASO 1)
- Número de nombres que se encuentran en QGIS y no en R: 3 (CASO 2)
- Número de nombres que son comúnes: 8 (CASO 3)

Para poder realizar una correcta explicación del proceso, enumeramos los casos como se puede ver anteriormente en el resumen de casos.

A continuación se explica lo que hacemos en cada caso:

Caso 1:

En este caso, para cada nombre habría que añadir una fila en QGIS únicamente rellenando el campo de nombre ya que los demás ya se van a rellenar con los datos de R. También habría que buscar informacion de la localización de esta estación y rellenar los campos necesarios para geolocalizar esta estación en QGIS.

Caso 2:

En este caso, la medida que emplearemos para "intentar conocer" los valores de contaminación en el aire de estas estaciones será la interpolación, es decir a partir de los datos de las estaciones próximas a los puntos con información faltante (Dr.Lluch, Patraix y Cabanyal), se podrá calcular una aproximación de los valores de contaminación para esas estaciones.

Caso 3

Este caso es el más simple de todos ya que lo único que habrá que hacer será importar los datos a QGIS y hacerlos coincidir mediante el campo "Estacion" que se encuentra tanto en el conjunto de datos de R y el conjunto de puntos de QGIS.

Objetivo de la Exportación a QGIS:

El objetivo de la exportación a QGIS de los datos es poder realizar un analisis espacial de los datos teniendo en cuenta los valores de las distintas partículas en suspensión que se encuentran en el aire.

Procedimiento:

En primer lugar, creamos un data frame solo con las estaciones que son comúnes a R y QGIS: CASO3.

```
comunes <- df_num %>%
  filter(Estacion != "Puerto llit antic Turia(s)") %>%
  filter(Estacion != "Puerto Moll Trans. Ponent(s)")
```

Una vez hecho, exportamos para introducirlo en QGIS:

```
write.csv(comunes, file = "comunes.csv", row.names = FALSE)
```

Ahora cogemos las Estaciones correspondientes al CASO1 y hacemos un nuevo data frame con ellas, al que añadiremos dos nuevas variables que contendrán la longitud y latitud en el EPSG:4326.

```
CASO1 <- df_num %>%
  filter(Estacion == "Puerto llit antic Turia(s)" |
           Estacion == "Puerto Moll Trans. Ponent(s)") %>%
  mutate(
    x = ifelse(Estacion == "Puerto llit antic Turia(s)", -0.328992 ,
                     ifelse(Estacion == "Puerto Moll Trans. Ponent(s)", -0.323220, NA))) %>%
  mutate(
    y = ifelse(Estacion == "Puerto llit antic Turia(s)", 39.450454 ,
                     ifelse(Estacion == "Puerto Moll Trans. Ponent(s)", 39.459204, NA)))
```

Una vez hecho, se exporta para introducirlo en QGIS:

```
write.csv(CASO1, file = "CASO1.csv", row.names = FALSE)
```

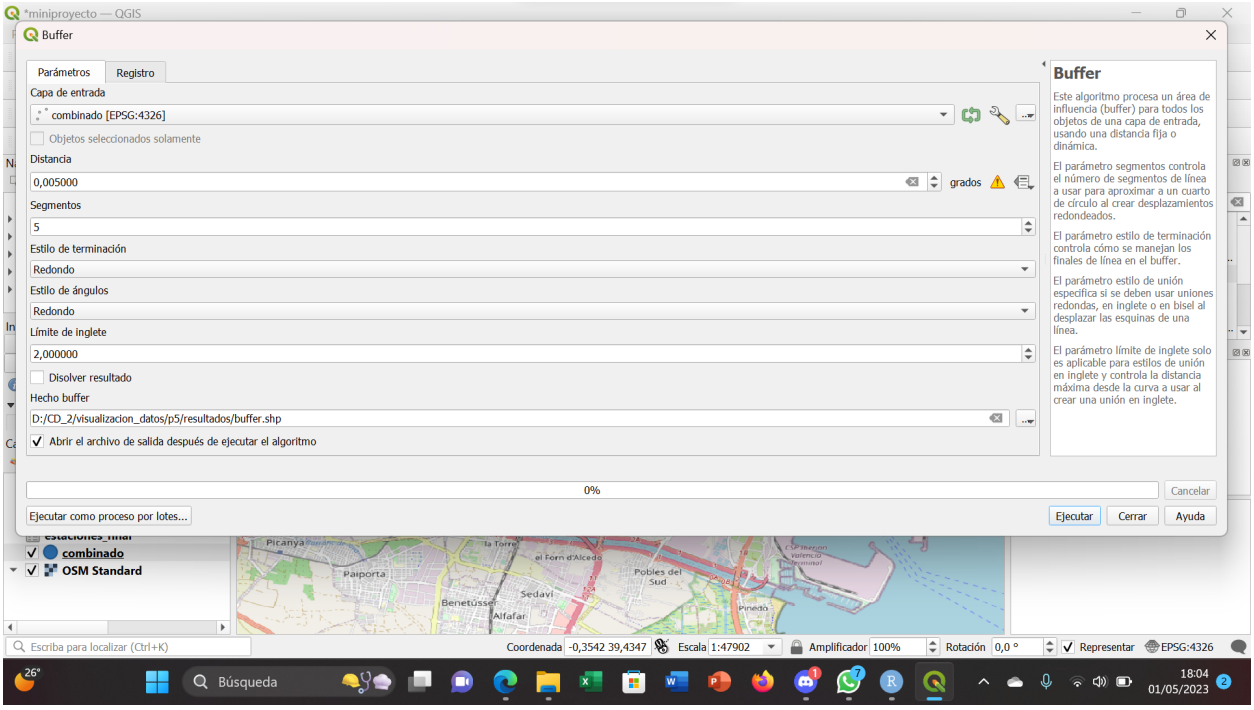
Introducimos a QGIS como capa de texto delimitado CASO1 y seleccionamos los campos x e y como coordenadas. En la capa estacions-contaminacio, realizamos la interpolación de las tres estaciones Dr.Lluch, Patraix y Cabanyal para obtener de forma aproximada los valores faltantes.

Una vez tenemos las capas CASO1 y estacions-contaminacio, modificamos los campos de cada una de las capas para que tengan el mismo nombre y el mismo tipo de dato.

Como los únicos campos que no coinciden entre ambas capas son por parte de la capa “CASO1” las variables x e y, por otro lado, en la capa “estacions-contaminacio” la variable globalid, al unir estas capas, obtenemos una capa final con todas las estaciones y sus respectivos datos sin importar qué variables usan como coordenadas. Es decir, los atributos de esta nueva capa son unas estaciones que tienen valores en las columnas X e Y y valores NULL en globalid, y otras estaciones con valores en globalid y no en X e Y.

6. Buffer

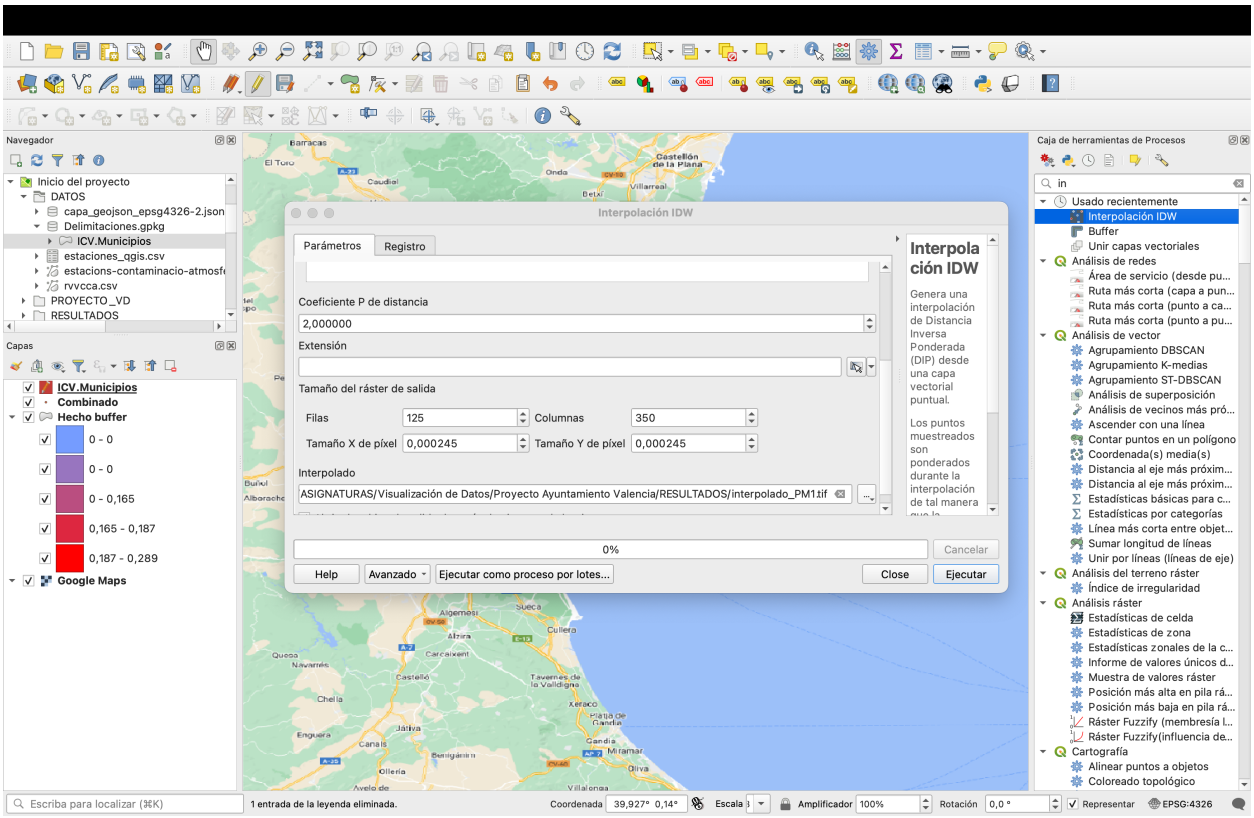
Realizamos una capa de áreas de influencia para cada estación.



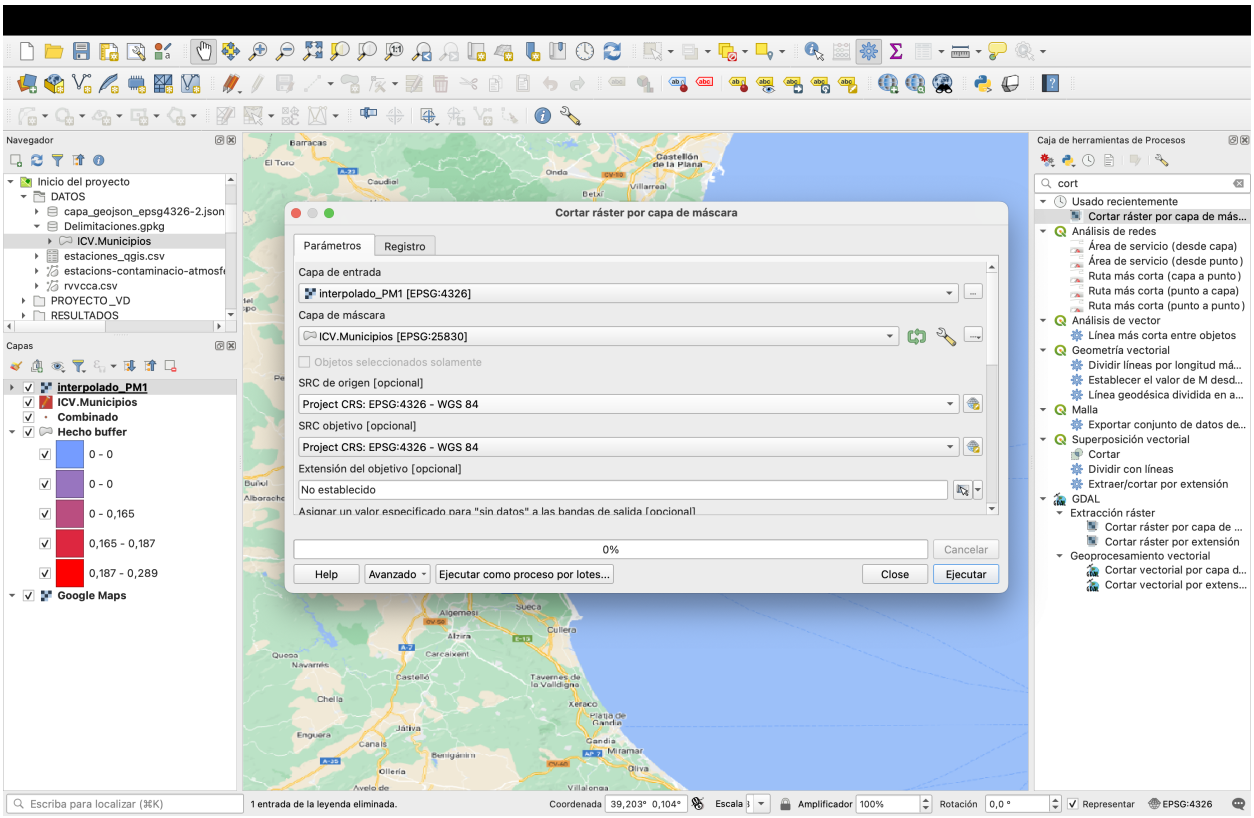
7. Interpolación

Hacemos una interpolación IDW para cada sustancia química que se puede encontrar en cada estación. Para ello, se estiman valores en ubicaciones desconocidas con el fin de crear una superficie ráster que cubra un área completa

Las siguientes imágenes muestran únicamente cómo se realiza para una sustancia química.



Sin embargo, obtenemos un ráster rectangular, ya que los puntos de muestreo se ponderan durante la interpolación de tal manera que la influencia de un punto en relación con otros disminuye con la distancia desde el punto desconocido que se desea crear. El ráster debemos cortarlo por capa de máscara para visualizar únicamente las zonas de Valencia que necesitamos



8. Aplicación

Librerías de la app

```
#library(shiny)
#library(ggplot2)
#library(DT)
#library(shinythemes)
#library(dplyr)
#library(readr)
#library(leaflet)
#library(leaflet.opacity)
#library(sf)
#library(raster)
#library(tidyr)
#library(ggthemes)
#library(plotly)
```

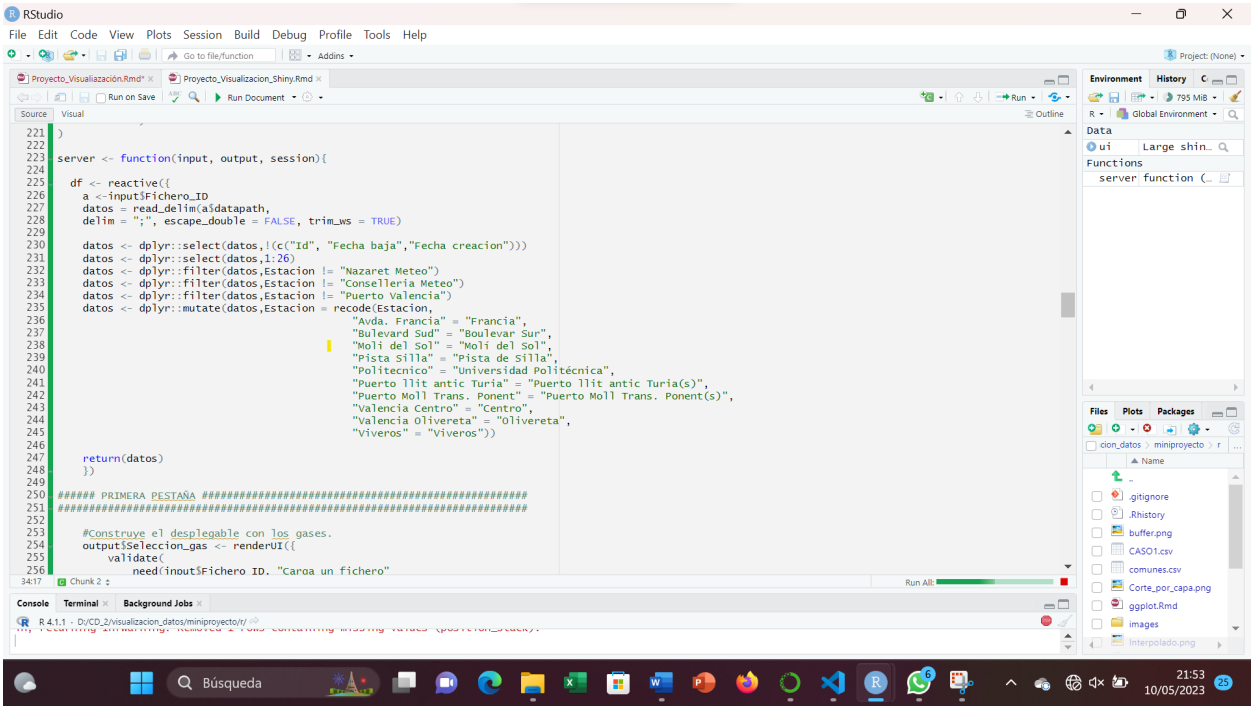
Mediante shiny, realizamos una aplicación que contendrá tres menús:

1. El menú principal indica un resumen del funcionamiento de la aplicación. Además, permite cargar el fichero de datos a usar.

2. Menú desplegable en el que se realizan distintos gráficos que el usuario podrá elegir visualizar.

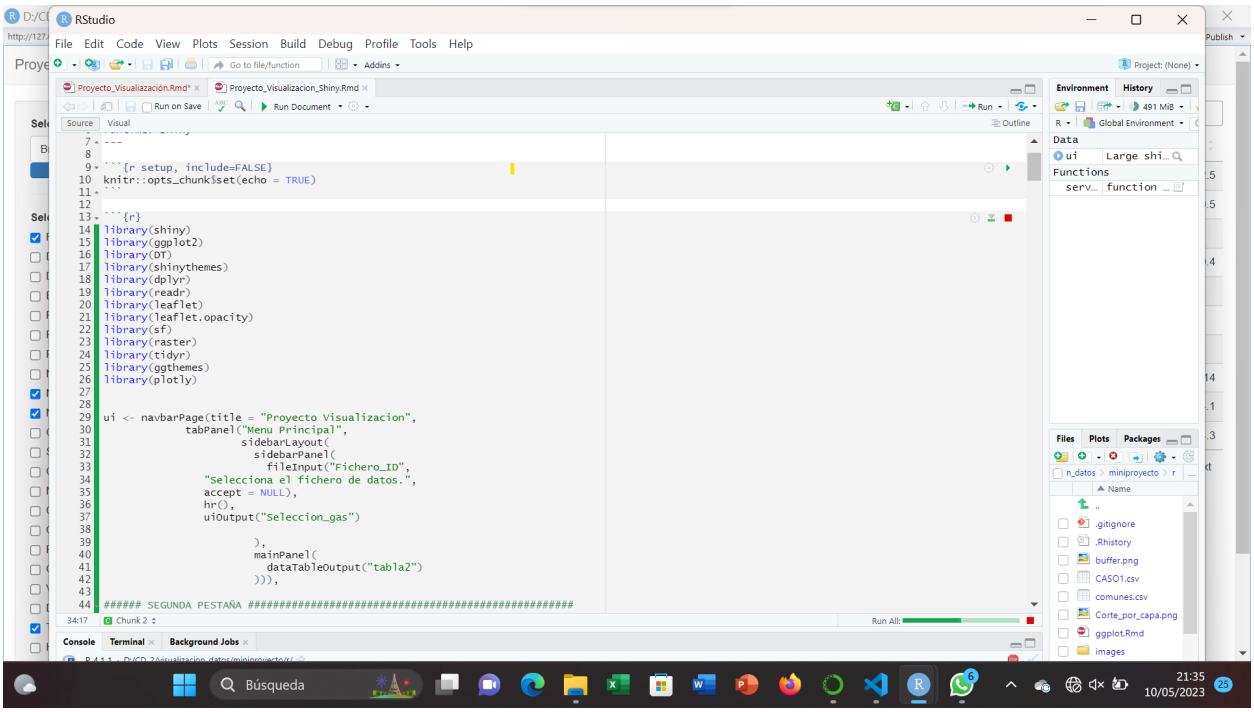
3. Menú para una visualización espacial general.

Cargamos todas las librerias necesarias para la ejecución del código y definimos la UI y el servidor. En el servidor, creamos un objeto reactivo correspondiente al archivo que se introducirá y que se modifica eligiendo las filas/columnas que necesitamos y cambiando los nombres de las estaciones.

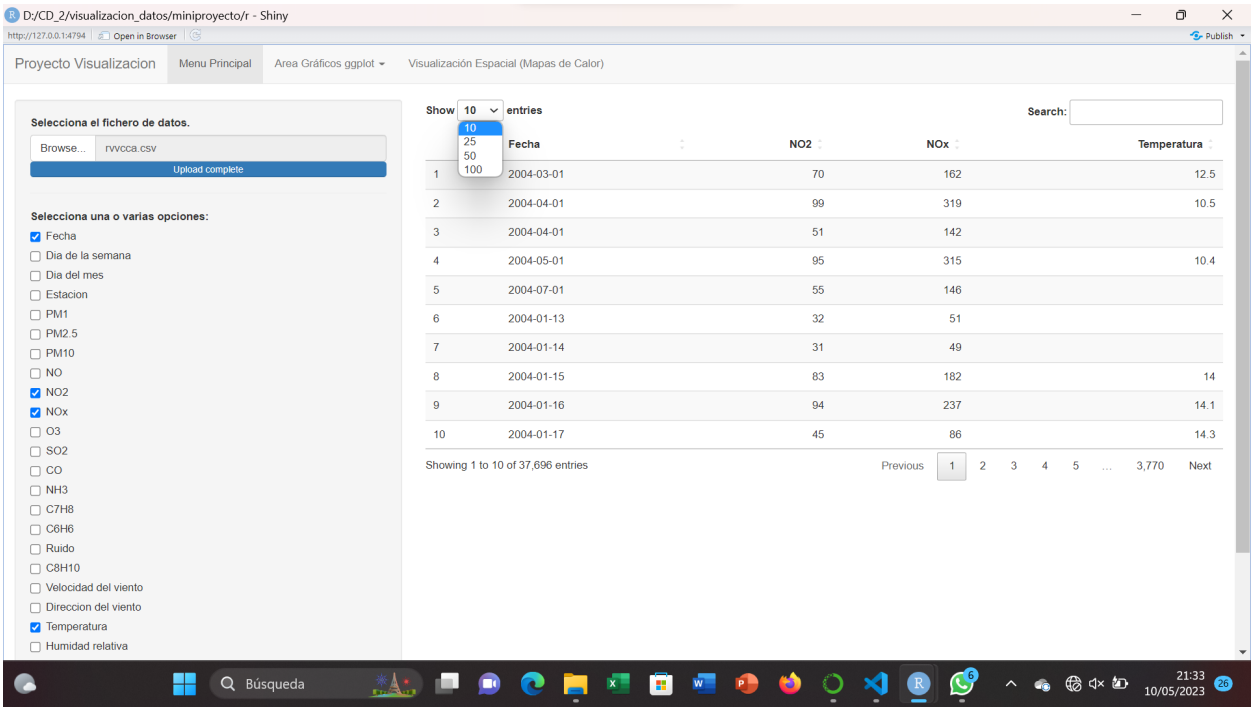


1. Menú principal

En el programa, este menú se ha diseñado mediante navbarPage.

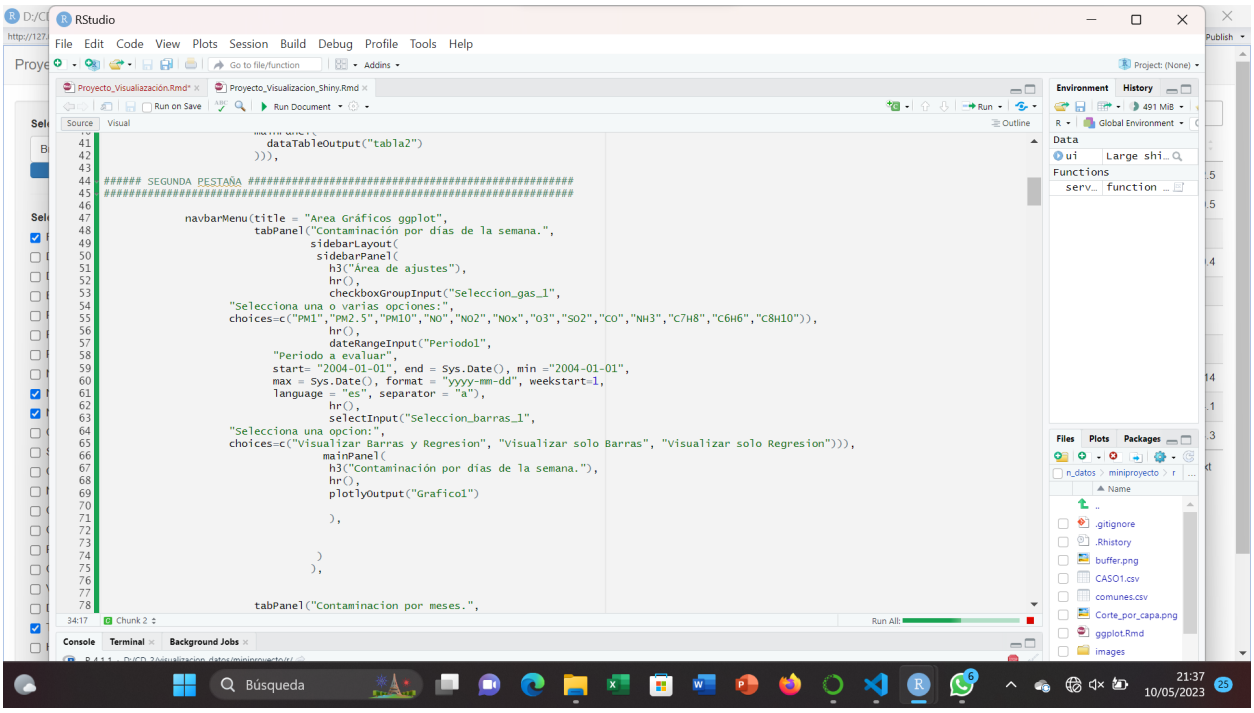


Cargamos el archivo rvvca.csv para que en los siguientes menús podamos realizar las diferentes visualizaciones. Podemos ver un breve contenido (en forma de tabla) del archivo con el número de entradas que indiquemos, así como buscar un dato que se quiera.



2. Área Gráficos Ggplot

Como es un menú desplegable realizado mediante navbarMenu, podremos elegir los distintos gráficos hechos en la UI con tabPanel:

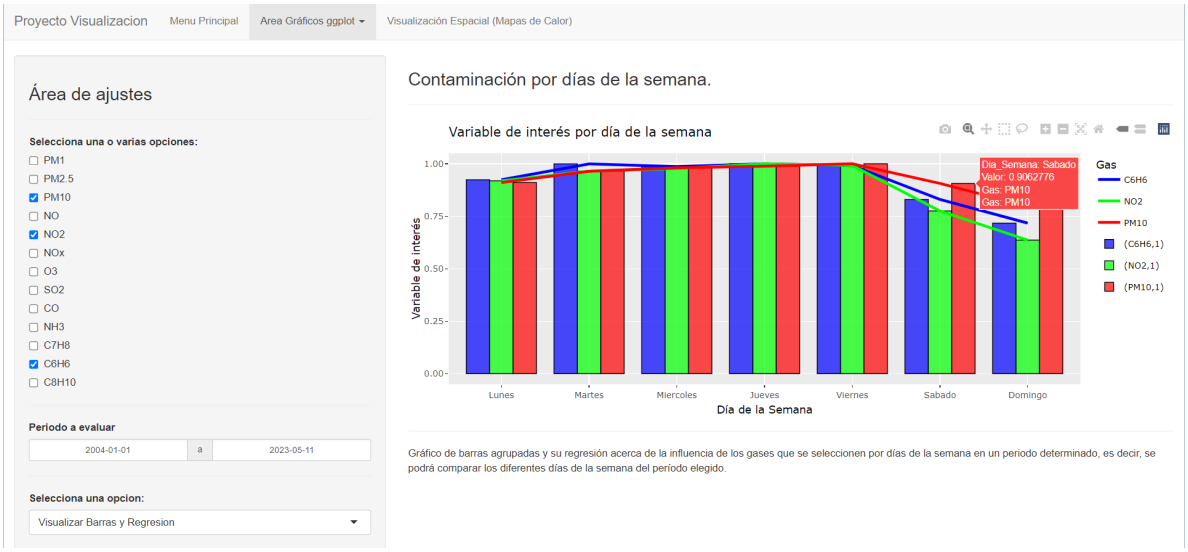


- Contaminación por días de la semana:

  Tenemos una gráfica en la cual podemos elegir los diferentes gases que se visualizarán de diferente color durante un período determinado.

  Los valores de los gases será la variable respuesta representados en el eje Y; en el eje X se muestran los gases seleccionados de diferente color para que se visualicen mejor en cada día de la semana.

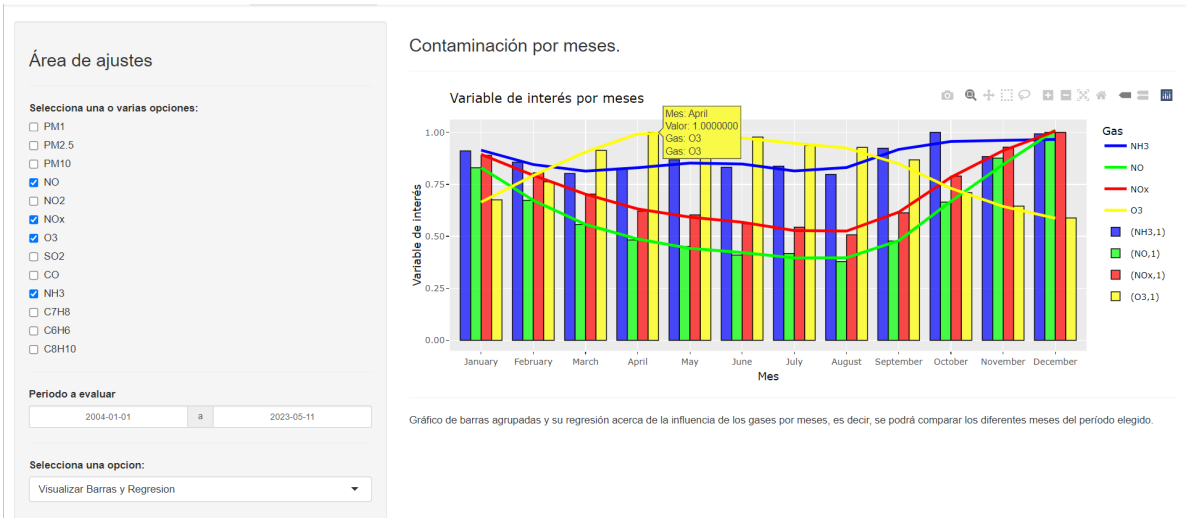
  Es posible ver las barras agrupadas y/o su regresión durante la semana.



- Contaminación por meses

  Tenemos una gráfica de barras agrupadas, donde podemos elegir los diferentes gases que se visualizarán de diferente color durante un período determinado. Como se superponen los meses, cambiamos su orientación para que se puedan ver correctamente.

  Los valores del gas será la variable respuesta representados en el eje Y; en el eje X se muestran los gases seleccionadas y los meses del año. Es posible visualizar las barras junto a la regresión que sigue, o únicamente la regresión.



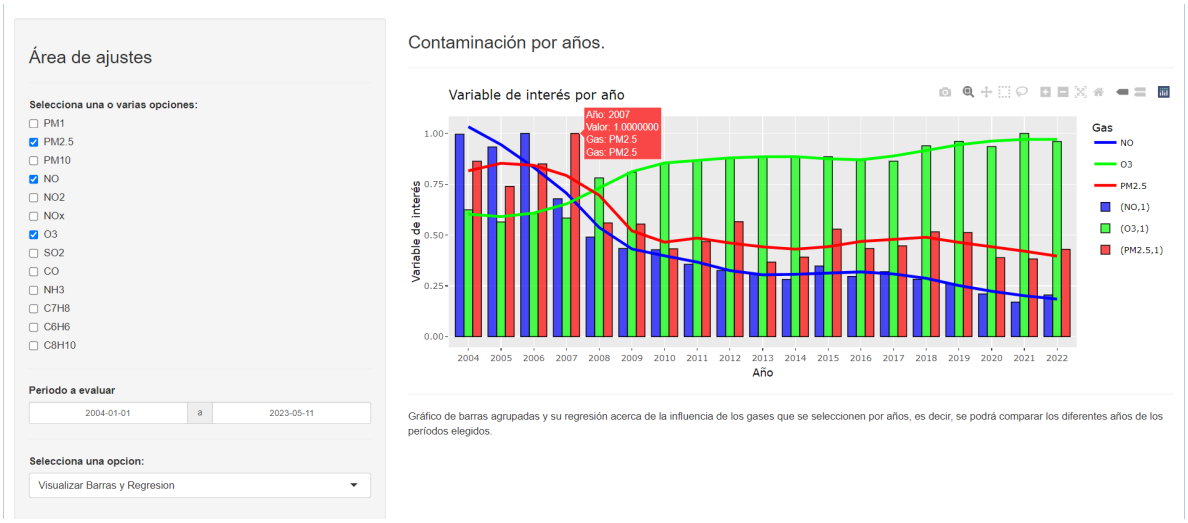
- Contaminación por años

  Obtenemos otra gráfica de barras agrupadas por los gases que se podrán seleccionar en un intervalo de tiempo determinado.

Los valores del gas será la variable respuesta representados en el eje Y; en el eje X se muestran los gases y los años seleccionados (se giran ya que es posible que se superpongan).

Se puede elegir si se quiere solo las barras, su regresión, o ambas.

En esta gráfica y las anteriores no se reorganizan las barras ni se intercambian los ejes ya que la figura resultante sería más confusa ya que podemos tratar varios gases y el tiempo es una variable que mantiene un orden natural. Además, hemos normalizado los datos (intervalo entre 0 y 1) para que puedan ser comparables.

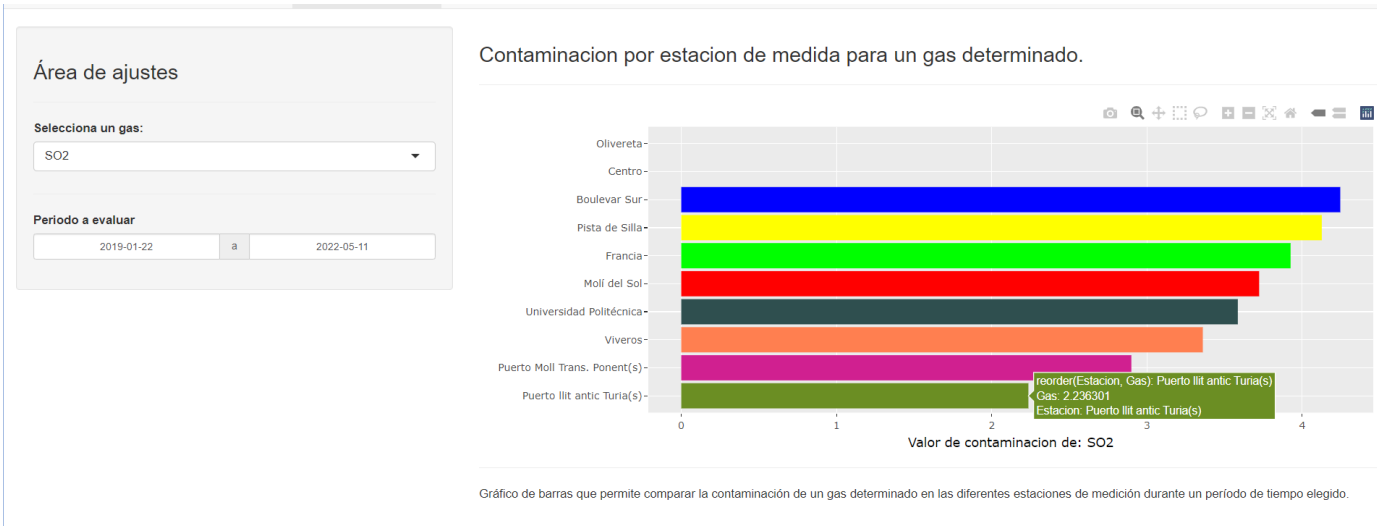


- Contaminación por estación de medida

  Tenemos una gráfica de barras de cada estación de medida en el que podemos elegir un período a evaluar y el valor de contaminación de un gas que se elija.

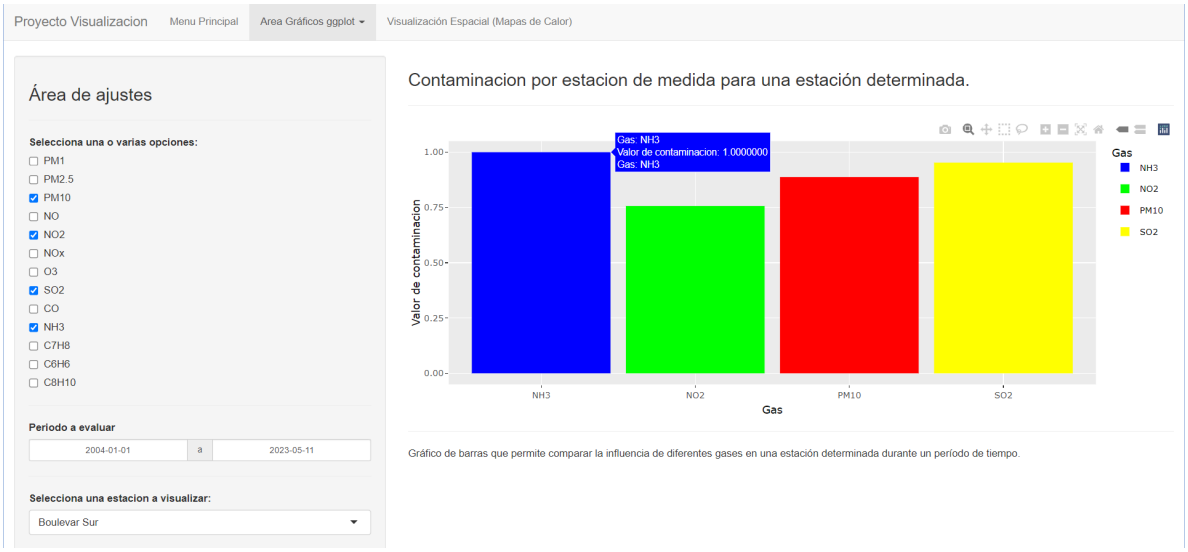
  Al principio, el eje x contenía las estaciones y el eje y el valor del gas seleccionado, sin embargo, se intercambiaron los ejes ya que así se permite una mejor visualización de los datos. Además, reorganizamos las barras dispuestas en orden de su tamaño para que sea más intuitiva ya que las categorías que representan las barras (las estaciones) no siguen un orden natural.

  En algunos gases, no tenemos información acerca de alguna estación de medida ya que no hay registros en el data frame.



- Contaminación e influencia de las precipitaciones

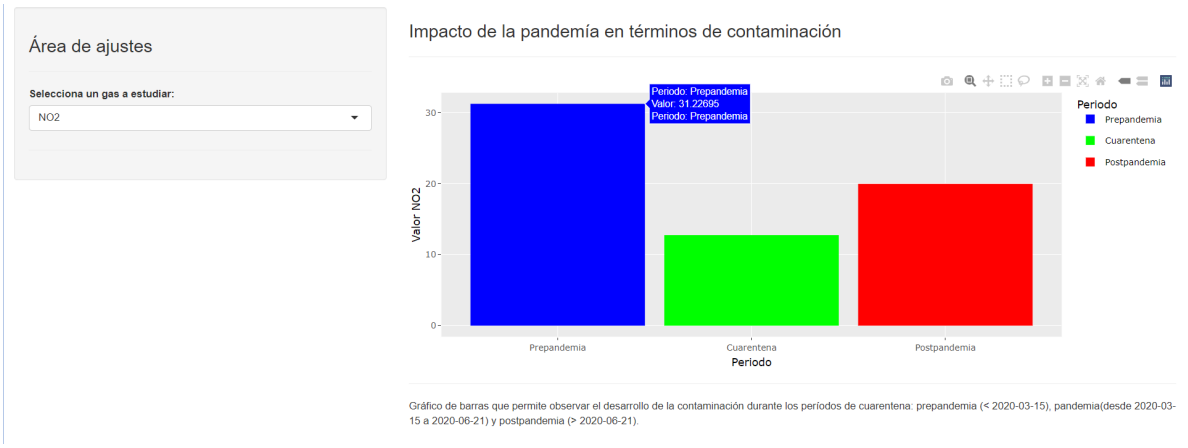
Al contrario que el gráfico anterior, en este caso se eligen diferentes gases para visualizar la contaminación que afecta en únicamente una estación durante el período que se elija. Cada gas se podrá comparar con los otros seleccionados ya que se rellenan de colores diferentes (una paleta de colores distintos ya que son variables categóricas que no tienen que ver entre sí, es decir, un gas es independiente de los demás) y porque hemos normalizado los datos (valores entre 0 y 1).



- Impacto de la pandemia en términos de contaminación

  Para esta gráfica, se seleccionaron las fechas que correspondían con prepandemia (<2020-03-15), pandemia(desde 2020-03-15 a 2020-06-21) y postpandemia (>2020-06-21).

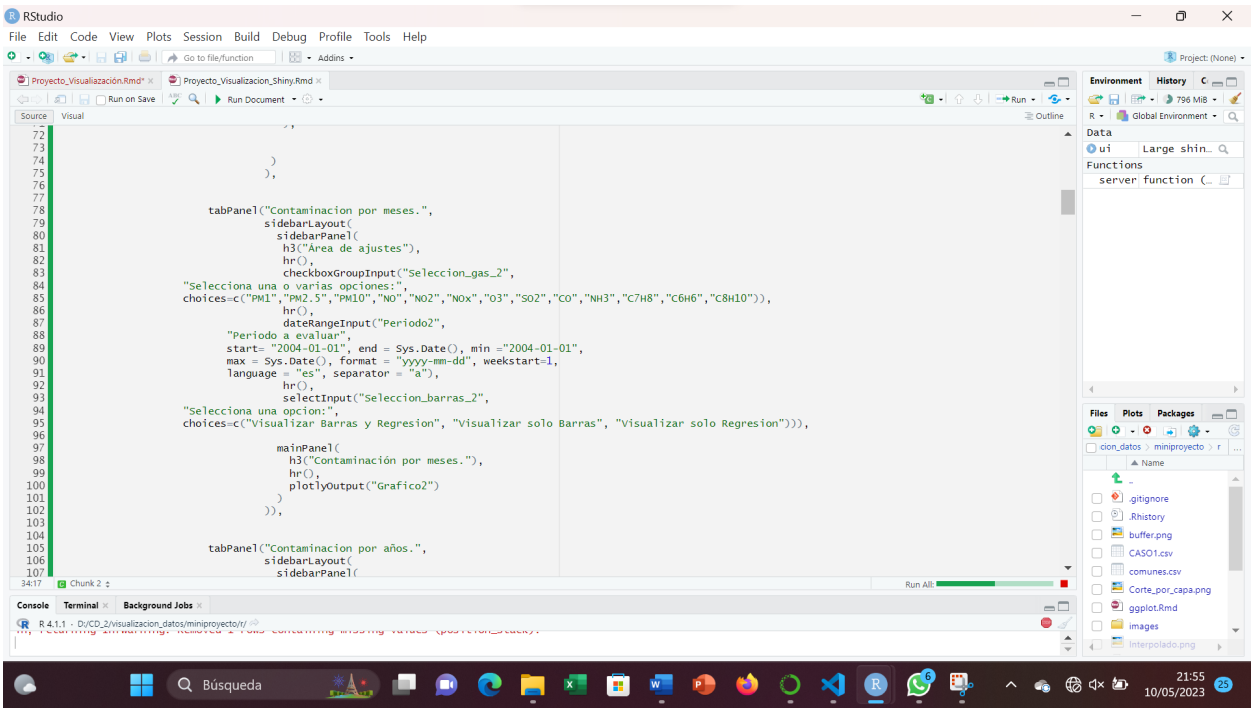
  En este caso, podemos ver la contaminación de un gas que se quiera para ver el cambio que ha habido durante los diferentes tiempos de pandemia.

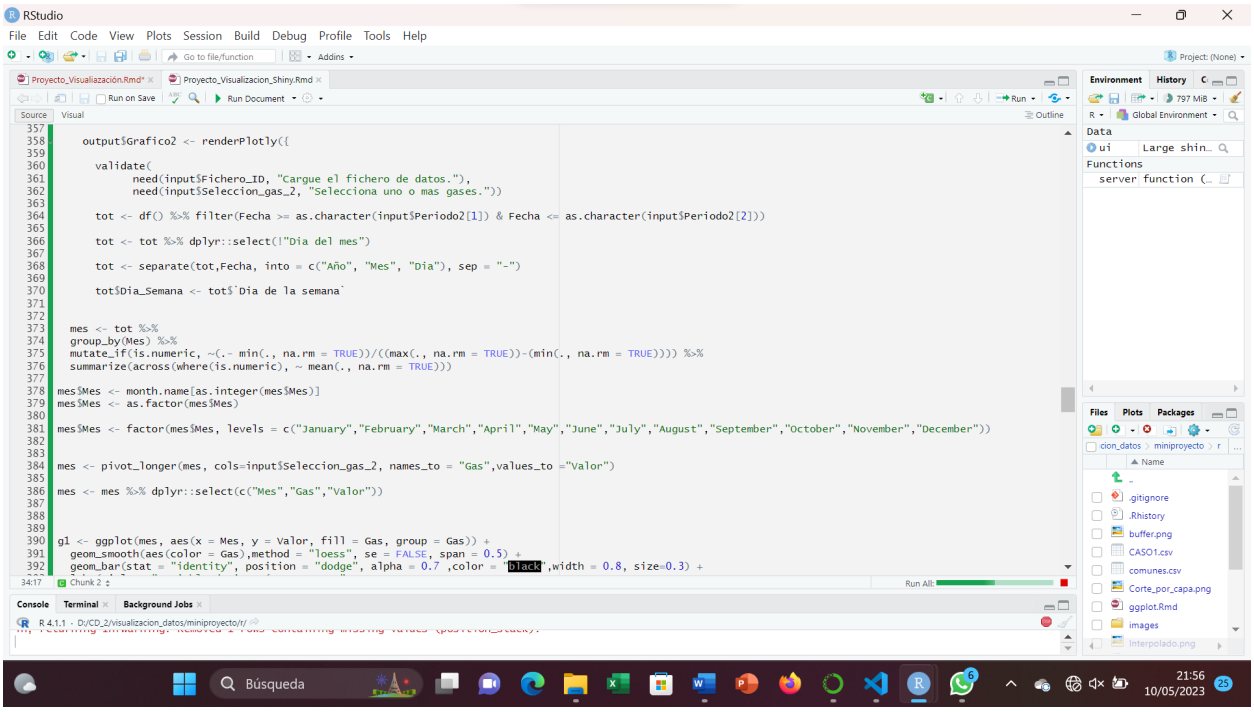


  Cada gráfico tiene la posibilidad de obtener los detalles (valor, nombre del gas, nombre de la estación...) si pasamos el cursor por cada dato gracias a la librería plotly.

El diseño de las páginas (texto, pestañas para opciones, períodos de tiempo...) se definen en la UI, mientras que los filtrados de los data frames y las representaciones que se utilizan en cada caso están en el servidor.

Aquí el ejemplo al segundo gráfico:





3. Visualización Espacial

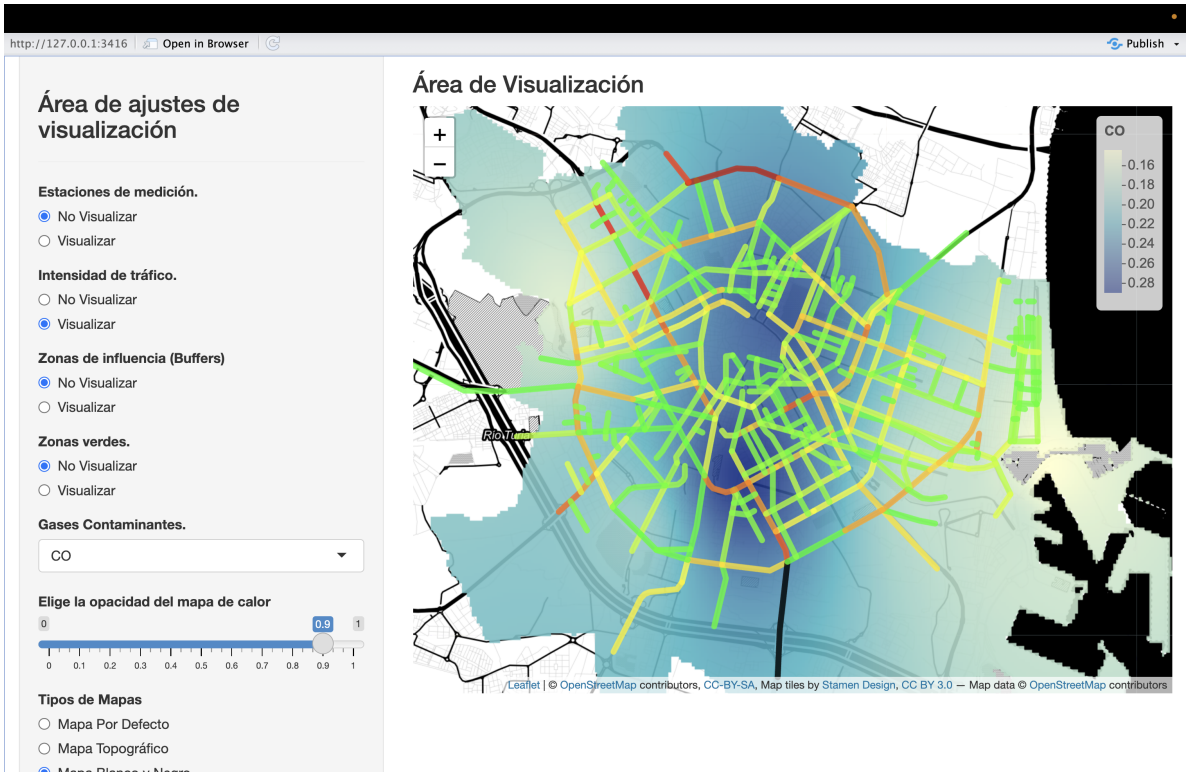
Esta página se ha diseñado mediante la implementacion de elementos interactivos ya que si se modifican las opciones de las entradas, se recalculará la visualización del mapa.

Mediante funciones de la librería de leaflet, se introduce un mapa de Valencia en el cual podemos hacer zoom para obtener detalles de algún lugar específico. Tenemos las opciones de:

- Visualizar las estaciones de medición (metidas como .shp de la capa final). Se indicarán con un marcador(del tamaño que se quiera), el cual se definió como un icono que descargamos previamente.
- Visualizar las zonas de influencia (buffer) de las diferentes estaciones introducidas como archivo .shp ya que se calcularon en QGIS.
- Visualizar la intensidad de tráfico. Se muestran las calles marcadas de un color que depende del número de coches que han pasado a lo largo del día. Con ello, se puede concluir la relación de que la emisión de gases de los coches será mayor contra más rojo esté; o por el contrario, estará más amarillo a menos contaminación haya.

  El archivo con formato .json se ha descargado de:

  http://datos.gob.es/es/catalogo/l01462508-intensidad-de-trafico-por-tramos

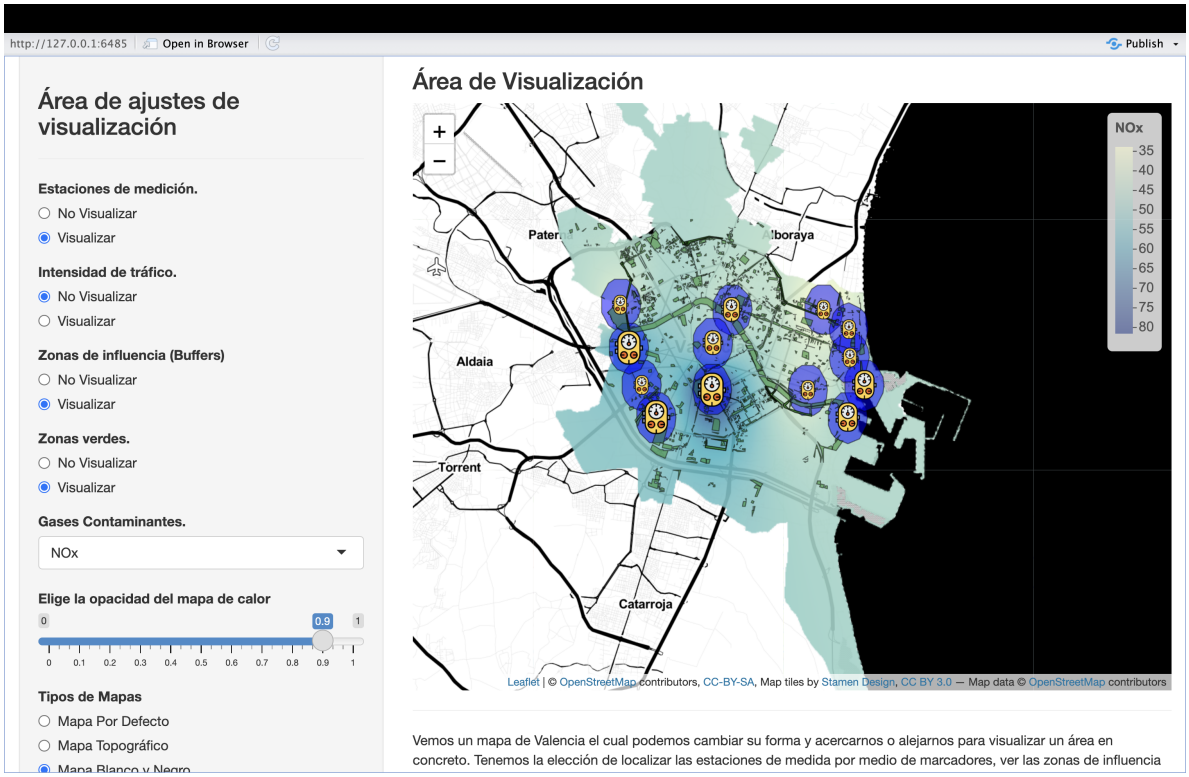


- Visualizar las zonas verdes obtenidas del data frame adicional para poder sacar otras conclusiones que puedan ser importantes.

  https://valencia.opendatasoft.com/explore/dataset/zonas-verdes/table/?disjunctive.nivel3

- Seleccionar un gas y obtener su interpolación (se añadieron como archivos .tif obtenidos por QGIS), pudiendo elegir la opacidad de esta.
- Tres diferentes tipos de mapas: El mapa por defecto de leaflet, un mapa topográfico o un mapa en blanco y negro.

  A continuación se muestra el mapa completo sin la intensidad del tráfico


http://127.0.0.1:6485
Open in Browser
Publish
Área de ajustes de visualización
Estaciones de medición.
No Visualizar
Visualizar
Intensidad de tráfico.
No Visualizar
Visualizar
Zonas de influencia (Buffers)
No Visualizar
Visualizar
Zonas verdes.
No Visualizar
Visualizar
Gases Contaminantes.
NOx
Elige la opacidad del mapa de calor
0
0.9
1
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Tipos de Mapas
Mapa Por Defecto
Mapa Topográfico
Mapa Blanco y Negro
Área de Visualización
NOx
35
40
45
50
55
60
65
70
75
80
Aldaia
Torrent
Catarroja
Leaflet | © OpenStreetMap contributors, CC-BY-SA, Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors
Vemos un mapa de Valencia el cual podemos cambiar su forma y acercarnos o alejarnos para visualizar un área en concreto. Tenemos la elección de localizar las estaciones de medida por medio de marcadores, ver las zonas de influencia