PRA1. Tipología y Ciclo de datos

Autores: Kilian Cañizares, Aleix Cortina.

1. Contexto

Actualmente se está viviendo una situación excepcional a escala global debido a la propagación del virus SARS-CoV-2 que causa la enfermedad COVID-19 que en un bajo porcentaje (<5%) puede requerir ingresos en Unidades de Cuidados Intensivos (UCI)[1]. Debido a su elevada propagación, el sistema sanitario español corre el peligro de colapsar, por lo que el Gobierno ha decretado medidas de confinamiento a la población con las consecuencias económicas, sociales y ambientales que esto conlleva. Este conjunto de datos debe de servir para evaluar las consecuencias ambientales de estas medidas.

2. Titulo

Evolución del índice de calidad del aire durante el confinamiento en las poblaciones con más habitantes de cada Comunidad Autónoma de España.

3. Descripción del conjunto de datos

Los datos son recolectados por estaciones de muestreo automáticas que efectúan mediciones de:

- PM10. Material particulado menor de 10 micrómetros
- O_3. Ozono
- NO_2. Dióxido de Nitrógeno.

Estos componentes pueden suponer un riesgo para la salud humana a medida que aumentan su concentración. Los valores recogidos por las estaciones de muestreo automáticas son en concentración (μg partícula/m^3 aire) y son convertidos a niveles de Índice de Calidad del Aire (AQI, Air Quality Index por sus siglas en inglés) a partir de estándares de la Agencia de Protección de Medio Ambiente (EPA, Environmental Protection Agency por sus siglas en inglés) de Estados Unidos[2]. Valores bajos de AQI son considerados más saludables para el ser humano, y valores altos más peligrosos, con la siguiente escala de peligrosidad:

Rango AQI	La calidad del aire es
0-50	Buena
51-100	Moderada
101-150	Insalubre para grupos sensibles
151-200	Insalubre
201-300	Muy insalubre
301-500	Peligrosa

En el caso de existir el histórico de datos, el rango de fechas recolectado pertenece a los años 2019-2020. De esta manera, se pretende poder observar el efecto del confinamiento decretado por el Gobierno en marzo de 2020 y la posible variabilidad interanual entre los dos años. Se ha obtenido el índice AQI para la ciudad más poblada de cada Comunidad Autónoma a excepción de las Comunidades Autónomas de Castilla la Mancha, Extremadura, Principado de Asturias, Ceuta, Melilla y La Rioja donde no existen mediciones. Para las ciudades de Madrid y Barcelona se han encontrado datos de agregados de todas las estaciones de monitorización, en cambio para el resto de las ciudades se ha escogido la estación de medición más cercana al centro de la ciudad.

Como aporte de material complementario, se adjunta un conjunto de datos obtenidos de la página del Instituto de Salud Carlos III, donde se detallan los casos de COVID-19 por Comunidad Autónoma, para que de esta manera los usuarios del juego de datos de calidad del aire tengan un marco de referencia de la evolución de la enfermedad en España.

En ambos casos se examinó el fichero “robots.txt”. En el caso de los datos de calidad atmosférica (https://waqi.info) se permite el acceso a todos los robots y a todos los ficheros. En el caso de los datos de casos de COVID-19 (https://www.isciii.es/), se permite el acceso a todos los robots y se excluyen los directorios: “/_layouts/”, “/_vti_bin/” y “/_catalogs/”. Se adjuntan los ficheros “robots.txt” en el directorio “/robot_file” del repositorio.

4. Representación gráfica

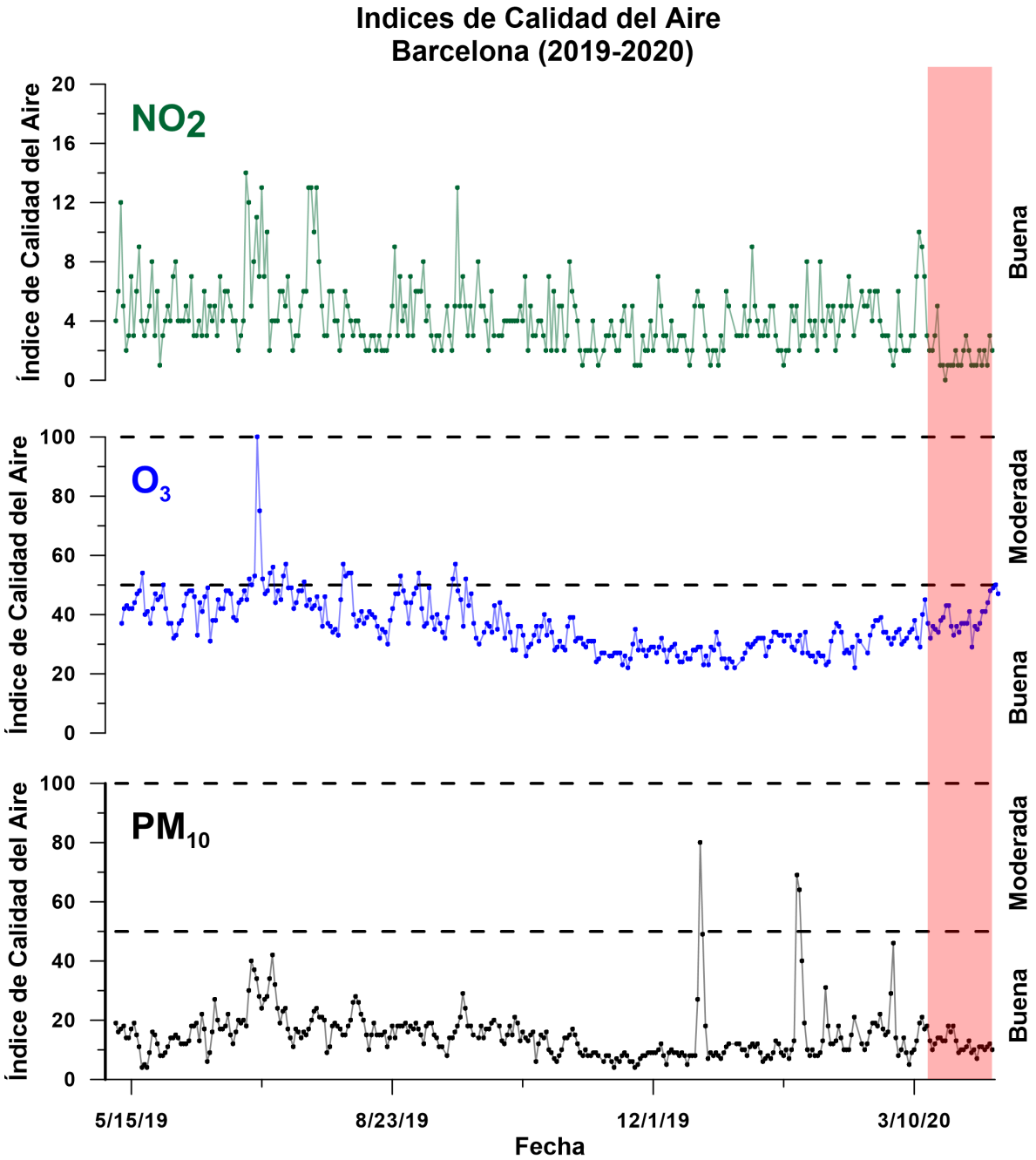

Figura1. Índices de calidad del aire para el periodo 2019-2020 de la ciudad de Barcelona. Se representan los índices AQI (Air Quality Index) de los compuestos PM_{10}, O_3 y NO_2. Las líneas punteadas horizontales separan AQI que consideran la calidad del aire buena (i.e. 0-50) o moderada (i.e. 51-100). La barra roja comprende el periodo de tiempo donde se ha decretado el confinamiento de la población por el Gobierno de España.

5. Contenido
Conjunto de datos referente a la calidad del aire hasta el día 9 de abril de 2020 presentes en el fichero "air_contamination.csv":

- **timestamp**. Fecha en formato "yyyy/mm/dd".
- **ca.** Comunidad Autónoma.
- **ciudad**. Ciudad más poblada de la Comunidad Autónoma.
- **pm10**. Valor AQI del material particulado menor de 10 micrómetros.
- **pm10_level.** Valor categórico que indica la calidad del aire en referencia al valor de PM_{10} según su valor AQI.
- **o3**. Valor AQI del Ozono.
- **o3_level.** Valor categórico que indica la calidad del aire en referencia al valor de O_3 según su valor AQI.
- **no2**. Valor AQI del Dióxido de Nitrógeno.
- **no2_level.** Valor categórico que indica la calidad del aire en referencia al valor de NO_2 según su valor AQI.

Para realizar la extracción de los datos de la página https://waqi.info se realizaron los siguientes pasos:

1. Se consulta el fichero "robots.txt" para comprobar las limitaciones.
2. Se consulta el fichero "sitemaps.xml" (la página no dispone de fichero "sitemap.xml").
3. Se analiza el código fuente de la página en busca de los datos objetivo que se habían identificado previamente. Al analizar el código se observa que los datos no están siempre presentes en la página, ya que es un script que se ejecuta cuando el usuario interactúa con la página.
4. Se implementa un "robot" que imitará el comportamiento de un humano para interactuar con la página y evaluar el script. Para ello, se ha utilizado la librería selenium para Python[3]. Gracias a esta librería y realizando un análisis del código fuente de la web es posible realizar un programa que interactua con la página web. El "robot" se dirige hacia el lugar donde se ejecuta el script que muestra los datos, y con el objeto de evitar colapsar el servidor de peticiones se comprueba la presencia de los datos cada 10 segundos.
5. Se descarga el código fuente de la página objetivo utilizando selenium, de esta forma, se asegura que los datos han sido evaluados.
6. Se crea una "soup" del HTML utilizando la librería beatifulsoup4 de Python[4].
7. Se busca en los elementos anidados de la sopa hasta encontrar la tabla dónde se encuentran los datos objetivo (i.e. tabla html con identificador "*historic-aqidata-inner*").
8. Se crea un diccionario con los datos objetivo para cada partícula de una ciudad.
9. Se crea el "dataframe" con los datos.
10. Se itere el proceso con las diferentes ciudades agregando los datos al "dataframe".
11. Finalmente, se guardan los datos del dataframe en un fichero ".csv".

Conjunto de datos referente a los casos de COVID-19 hasta el día 9 de abril de 2020 presentes en el fichero "casos_covid19.csv":

- comunidad. Comunidad Autónoma.
- casos. Número de casos de COVID-19 notificados por las Comunidades Autónomas al Ministerio de Sanidad.
- **casos_notificados**. Número de casos de COVID-19 notificados a la RENAVE (Red Nacional de Vigilancia Epidemiológica) a través de la plataforma SiVIES.
- **datetime**. Fecha en formato "yyyy/mm/dd".

Para realizar la extracción de los datos de la página https://www.isciii.es/ se han realizado los siguientes pasos:

1. Se consulta el fichero "robots.txt" para comprobar las limitaciones.
2. Se consulta el fichero "sitemaps.xml" (no contiene información relevante ni frecuencias de actualización).
3. Se analizan los datos objetivo, en este caso, los datos se encuentran en formato PDF, la página objetivo recopila los enlaces.
4. Se realiza un "get" a la página con la librería requests[5] de Python.
5. Se realiza una "soup" HTML con la librería beatifulsoup4[4] de Python.
6. Se buscan las etiquetas anidadas dónde se encuentran los enlaces a los scripts y se almacenan los enlaces de los PDFs que contienen los datos.
7. Se descarga cada PDF y se guarda en una carpeta temporal.
8. Se extrae la fecha de la publicación de los datos del nombre del PDF utilizando expresiones regulares con Python.
9. Se abre el PDF utilizando la librería tabula-py[6] de Python en un fichero JSON.
10. Se busca en el fichero JSON que contiene los datos del PDF de forma anidada los datos objetivo.
11. Se guardan los datos objetivo dentro de un diccionario de datos.
12. Se crea un "dataset" con los datos.
13. Se guardan los datos extraídos dentro de un fichero ".csv".

6. Agradecimientos

El acceso a los datos de calidad del aire ha sido gracias al proyecto "World Air Quality Index" cuyos datos están alojados en la web: https://waqi.info. También queremos agradecer al Instituto de Salud Carlos III por la recopilación de los datos de casos de COVID-19 por Comunidad Autónoma alojados en su página: https://www.isciii.es/.

7. Inspiración

El motivo para escoger este conjunto de datos está estrechamente relacionado con estudios de salud humana. Según la Organización Mundial de la Salud (OMS), alrededor de 4.2 millones de personas mueren al año debido a la contaminación ambiental[7]. De entre los diferentes compuestos presentes en las grandes ciudades, tienen especial importancia por

su peligrosidad para el ser humano: (1) la materia orgánica particulada, (2) el Ozono y (3), el Dióxido de Nitrógeno.

Por lo tanto, el estudio de la evolución de estos compuestos presentes en la atmósfera de las grandes ciudades es de vital importancia para el seguimiento e implementación de políticas destinadas a la mejora de la calidad de vida del ser humano. En estos momentos, y debido a la paralización de la actividad económica y a la reducción de la movilidad como resultado de las políticas de confinamiento aplicadas para la disminución de la propagación del virus SARS-CoV-2, se esperan cambios en los índices de calidad del aire asociados a estos procesos.

Esta situación excepcional es una excelente oportunidad para evaluar la posible mejora en los índices de calidad atmosférica en las ciudades más pobladas del Estado. Sin embargo, el acceso de los datos es difícil ya que cada administración autonómica los aloja en su servidor, y no en todos los casos hay índices de calidad asociados a ellos. El proyecto "World Air Quality Index" engloba todas las regiones a nivel mundial, y calcula el índice AQI para los diferentes compuestos. Aunque el acceso a los datos en tiempo real es sencillo a través de la API, los datos históricos han de ser descargados ciudad por ciudad en ficheros separados con el consiguiente consumo de tiempo y la dificultad de descargar datos a diferentes escalas geográficas.

Este juego de datos da, por lo tanto, respuesta a las necesidades de organizaciones tanto publicas como privadas que quieran evaluar la evolución de la calidad del aire durante esta situación de excepcionalidad en diferentes escalas geográficas.

8. Licencia

La licencia escogida es:

- **Released Under CC0: Public Domain License**. Se renuncia a los derechos de la obra bajo las leyes de derechos autorales en todo el mundo. Se puede copiar, modificar, distribuir e interpretar el juego de datos incluso para propósitos comerciales sin pedir permiso.

Se ha escogido este tipo de licencia ya que los datos ambientales son extraídos del proyecto "Air Quality Index" que es un proyecto sin ánimo de lucro y colaborativo que intenta incentivar la preocupación de la sociedad por una mejora en la calidad del aire. Este proyecto no recibe fondos públicos y sus ingresos para el mantenimiento del sitio vienen de la publicidad insertada en la página. Por ello y en consonancia con el propósito del proyecto se opta por la licencia arriba especificada. Además, los datos complementarios de casos de COVID-19 están recopilados del Instituto de Salud Carlos III, que es un ente público sin ánimo de lucro que recibe fondos estatales.

9. Código fuente

El código con el que se ha generado el presente “dataset” está presente en GitHub a través del siguiente link:

“https://github.com/shiny-data-scientist/webscrap_pract_1”

10. Dataset y DOI

Para citar todas las versiones del “DataSet” y los scripts usar:
DOI: 10.5281/zenodo.3748442.

11. Contribuciones

Contribuciones	Firma
Investigación previa	K.C., A.C.
Redacción de las respuestas	K.C., A.C.
Desarrollo código	K.C., A.C.

Referencias

1. https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/documentos/20200224.Preguntas_respuestas_COVID-19.pdf
2. https://www.epa.ie/air/quality/index/
3. https://selenium-python.readthedocs.io/
4. https://pypi.org/project/beautifulsoup4/
5. https://requests.readthedocs.io/en/master/
6. https://pypi.org/project/tabula-py/
7. http://www9.who.int/airpollution/en/