# INFORME DE DATA CURATION

# DE LA CIUDAD DE MADRID

## 1. Fuentes de datos

El proyecto estudia los datos recogidos por sensores distribuidos por las diferentes estaciones de control de la ciudad de Madrid, midiendo tanto la calidad del aire, como el estado del tráfico.

La descripción de las fuentes empleadas y de la estructura de los datos recabados es la siguiente:

### 1.1. Situación de estaciones de medición de la calidad del aire

Este conjunto de datos contiene información de las localizaciones, así como los tipos de sensores ubicados y su análisis.

**Fuente:**

https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=9e42c176313eb410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

**Estructura:**

| Nombre | Descripción |
|---|---|
| **NÚMERO** | Identificador de la estación |
| **ESTACIÓN** | Nombre de la estación |
| **DIRECCIÓN** | Ubicacion de la estación |
| **LONGITUD** | Longitud geográfica |
| **LATITUD** | Latitud geográfica |
| **ALTITUD** | Altitud geográfica |
| **TIPO ESTACIÓN** | Tipo de estación (Urbana de fondo, Urbana de tráfico y Suburbana) |
| **CONTAMINANTE MEDIDO** | Indicador de contaminante ($NO_2$, $SO_2$, CO, PM 10, PM 2.5, $O_3$, BTX, HC) |
| **SENSORES METEOROLÓGICOS** | Tipos de sensores (UV, VV, DV, TMP, HR, PRB, RS, LL) |

### 1.2. Mediciones de la calidad del aire

En este caso, el conjunto de datos ofrece los niveles de contaminación atmosférica diarios del municipio de Madrid.

La información se subdivide en función de la magnitud medida y las técnicas usadas para ello.

**Fuente:**

https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=aecb88a7e2b73410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default%3E

**Estructura:**

| Nombre | Descripción |
|---|---|
| **PROVINCIA** | Código de la provincia |
| **MUNICIPIO** | Código del municipio |
| **ESTACIÓN** | Código de la estación |
| **MAGNITUD** | Magnitud de medida ($SO_2$, CO, NO, $NO_2$, PM 2.5, PM 10, Nox, $O_3$, TOL, BEN, EBE, MXY, PXY, OXY, TCH, $CH_4$, NM HC) |
| **PUNTO_MUESTREO** | Punto de recogida de la muestra: Recoge 3 argumentos separados por '_': Código del punto, de la magnitud y de la técnica de medida |
| **ANO** | Año medida |
| **MES** | Mes medida |
| **D0i** | Contaminación del día i |
| **V0i** | Validación de la contaminación. Solo validos los registros con 'V' |

## 1.3. Tramos de carretera

Este tercer conjunto de datos muestra la localización y datos básicos de los puntos de medida del tráfico.

**Fuente:**

https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=ee941ce6ba6d3410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

**Estructura:**

| Nombre | Descripción |
|---|---|
| **Tipo_elem** | Describe la tecnología del punto de medida: URB (tráfico urbano) o M-30 (tráfico interurbano) |
| **Distrito** | Código del distrito |
| **Id** | Identificador del punto de medida |
| **Cod_cent** | Código de centralización en los sistemas y que se corresponde con el campo <código> de otros conjuntos de datos como el de intensidad del tráfico en tiempo real |
| **Nombre** | Nombre de la ubicación del punto de medida |
| **Umt_x** | Coordenada del centroide de la representación del polígono del punto de medida |
| **Umt_y** | Coordenada del centroide de la representación del polígono del punto de medida |
| **Longitud** | Longitud geográfica |
| **Latitud** | Latitud geográfica |

### 1.4. Mediciones del estado del tráfico

En este caso, el conjunto de datos ofrece información sobre el control del tráfico de Madrid, a través de los detectores de vehículos en los puntos de medida. La base de datos SICTRAM los registra e integra en periodos de 15 minutos.

**Fuente:**

https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=33cb30c367e78410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

**Estructura:**

| Nombre | Descripción |
|---|---|
| **Id** | Identificador único del punto de medida |
| **Fecha** | Fecha con formato yyyy-mm-dd hh:mm:ss |
| **Tipo_elem** | Nombre del tipo de punto de medida: Urbano o M30 |
| **Intensidad** | Intensidad del punto de medida en el periodo de 15 minutos (vehículos/hora). Un valor negativo implica la ausencia de datos. |
| **Ocupación** | Tiempo de ocupación del punto de medida en el periodo de 15 minutos (%). Un valor negativo implica la ausencia de datos. |
| **Carga** | Carga de vehículos en el periodo de 15 minutos. Parámetro que tienen en cuenta intensidad, ocupación y capacidad de la vía y establece el grado de uso de la vía de 0 a 100. Un valor negativo implica la ausencia de datos. |
| **Vmed** | Velocidad media de los vehículos en el periodo de 15 minutos (Km/h). Sólo para puntos de medida interurbanos M30. Un valor negativo implica la ausencia de datos. |
| **Error** | Indicación de si ha habido al menos una muestra errónea o sustituida en el periodo de 15 minutos.<br>N: no ha habido errores ni sustituciones.<br>E: los parámetros de calidad de algunas de las muestras integradas no son óptimos.<br>S: Alguna de las muestras recibidas era totalmente errónea y no se ha integrado. |
| **Periodo_integración** | Número de muestras recibidas y consideradas para el periodo de integración. |

## 2. Proceso de curación de los datos

A continuación, se listan los pasos llevados a cabo en la curación de los datos descritos anteriormente y su código puede obtenerse en el Jupyter Notebook dataCurationMadrid.ipybn al que redirige el siguiente enlace:

URL

**Pasos realizados durante la curación de los datos**

1. Inicialmente se han imporatado las librerías necesarias para la realización del trabajo y se han descargado ambos archivos (.csv) usando la librería Pandas, en Python.

   (Pasos específicos realizados en la curación de los datos de la calidad del aire)

2. Se consideran únicamente válidos los datos recogidos con 'V' como valor de validación. Por lo tanto, la contaminación de los días que no tengan este valor se les ha asignado NaN.

3. En el formato original las fechas y las validaciones se presentan divididas por columnas, con frecuencia diaria, lo que dificulta la extracción de información, por ello, se ha creado una sola columna, date, con las fechas completas.
4. Así mismo, los valores de la magnitud del parámetro de contaminación medido con cada una de las técnicas utilizadas ( en cada uno de los puntos de muestreo) han sido incorporados en la columna CONTAMINACION_AIRE.
5. Para cada punto de muestreo se tienen 17 magnitudes de contaminación distintas y otras 17 técnicas de medida. Ante tal variedad se ha optado por un parámetro global que mida la media de las contaminaciones en cada punto de muestreo.
6. Este dataset contiene los datos diarios del año 2018, mientras que el dataset de tráfico se restringe a octubre del mismo año. Por lo tanto, el estu dio se realiza durante este mes, de forma que se obrenga concordancia entre los resultados. Por las mismas razones se ha seleccionado Moratalaz como punto de muestreo, ya que, además de Vallecas, es el único punto de muestreo presente en ambos conjuntos de datos. Tras estas restricciones no se observan valores NaN.

   (Pasos específicos realizados en la curación de los datos del estado de tráfico)
7. En este conjunto de datos la información viene recogida cada 15 minutos, por lo que se ha agrupado por días para la comparabilidad con el dataset anterior, tomando la media, de las medidas de tráfico que se ofrecen.
8. De nuevo, tomando el identificador de Moratalaz se eliminan todos los NaN. La columna error tiene N en todos sus valores, indicando que todos los datos han sido recogidos sin fallos ni sustituciones.

# 3. Análisis de los resultados

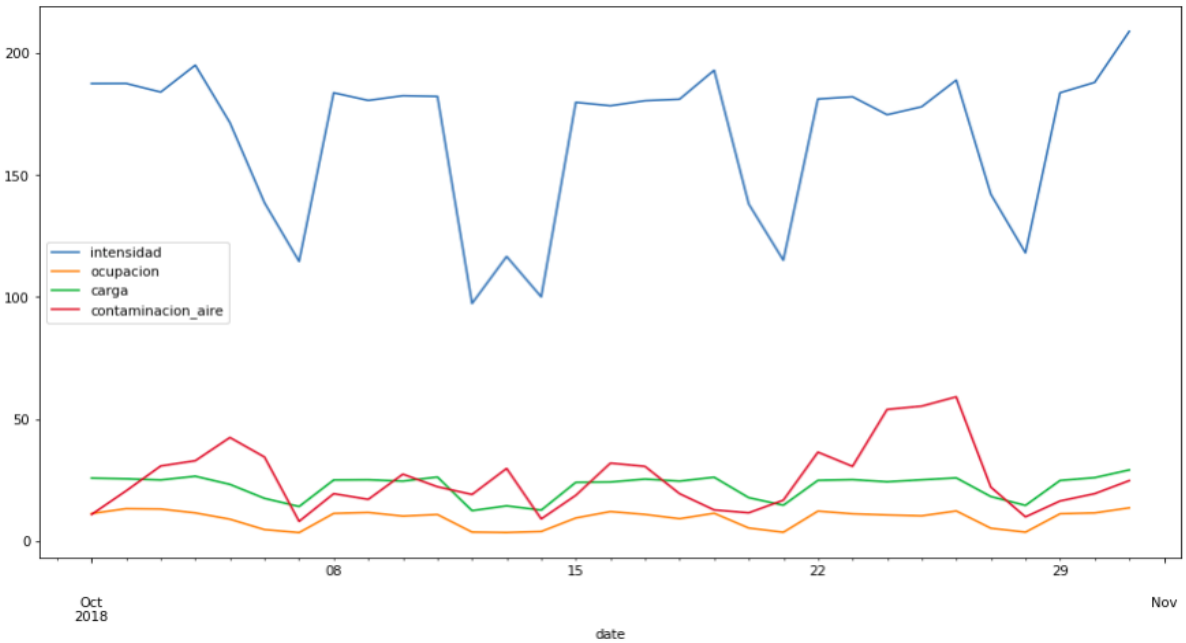


*Ilustración 1. Comparación contaminación y estado del tráfico de octubre en Madrid.*

En la representación de los resultados, no se ha podido observar una gran relación entre el estado del tráfico diario y la contaminación de ese día. Además, se ha realizado la matriz de correlación y lo confirma:

| | intensidad | ocupacion | carga | contaminacion_aire |
|---|---|---|---|---|
| **intensidad** | 1.000000 | 0.954977 | 0.995560 | 0.339095 |
| **ocupacion** | 0.954977 | 1.000000 | 0.962420 | 0.353054 |
| **carga** | 0.995560 | 0.962420 | 1.000000 | 0.356648 |
| **contaminacion_aire** | 0.339095 | 0.353054 | 0.356648 | 1.000000 |

*Ilustración 2. Matriz de correlación entre distintos factores de tráfico y la contaminación de Madrid según la Ilustración 1.*

Sin embargo, se ha realizado un análisis del estado del tráfico de días anteriores sobre la calidad del aire de cada día y se puede observar que la calidad del aire se ve afectada por la ocupación del tráfico de los tres días anteriores con una correlación de 0.53.

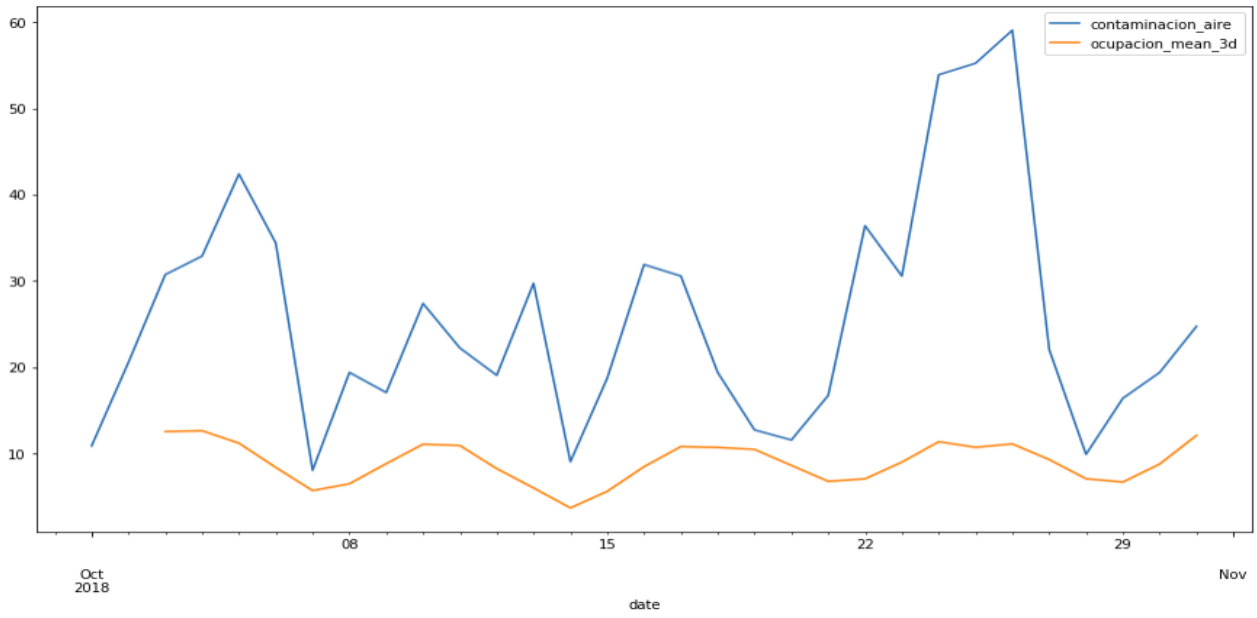


*Ilustración 3. Comparación contaminación del aire con la media de la ocupación del tráfico de los tres días anteriores.*