Informe Data Curation

Ciudad de Madrid

Datos utilizados

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 tanto de las fuentes empleadas como de la estructura de los datos recabados es la siguiente:

Situación de estaciones de medición de 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	ID de la estación
ESTACIÓN	Nombre de la estación
DIRECCIÓN	Ubicación 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 (NO2, SO2, CO, PM10, PM2,5, O3, BTX, HC)
SENSORES METEOROLÓGICOS	Tipo de sensores (UV, VV, DV, TMP, HR, PRB, RS, LL)

Mediciones de calidad del aire

En este caso el conjunto de datos ofrece los niveles de contaminación atmosférica del municipio de Madrid. La información se subdivide en función de la magnitud medida y las técnica usadas para ello. La información recogida por las estaciones de control tiene una frecuencia diaria.

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 (SO2, CO, NO, NO2, PM2.5, PM10, Nox, O3, TOL, BEN, EBE, MXY, PXY, OXY, TCH, CH4, NM HC)**
PUNTO_MUESTREO	**Punto de recogida de la muestra: Recoge 3 argumentos separados por '_': Codigo del Punto, Código de la Magnitud y Código de la Técnica de medida utilizada**
ANO	**Año medida**
MES	**Mes medida**
D0i	**Contaminación del día i**
V0i	**Validación de la contaminación del día i. Tan sólo se consideran válidos los identificados como 'V'**

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	**ID 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**

Mediciones del estado del tráfico

El último 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 sobre periodos de 15 minutos.

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

Estructura

Nombre	Descripción
id	**Identificación única del Punto de Medida**
fecha	**Fecha con formato yyyy-mm-dd hh:mi: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.**
ocupacion	**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 tiene 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. N: no ha habido errores ni sustituciones E: los parámetros de calidad de alguna de las muestras integradas no son óptimos. S: alguna de las muestras recibidas era totalmente errónea y no se ha integrado.**
periodo_integracion	**Número de muestras recibidas y consideradas para el periodo de integración.**

Pasos en la curación

La curación de los conjuntos de datos anteriores se ha llevado a cabo en el Notebook de Jupyter adjuntado. Se ofrece a continuación un resumen de los pasos realizados:

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

Calidad del aire

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

2. 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 sóla columna (date) con las fechas completas.

3. Así mismo, los valores de la magnitud del parámetro de contaminación medido con cada cada una de las técnicas utilizadas (en cada uno de los punto de muestreo) han sido incorporados en la columna CONTAMINACION_AIRE.

4. 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.

5. Este dataset contiene los datos diarios del año 2018, mientras que el dataset de tráfico se restringe a Octubre del mismo año. Por ello que el estudio se realiza durante este mes, de forma que se obtenga concordancia entre los resultados. Por las mismas razones se ha seleccionado Moratalz 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 NaNs.

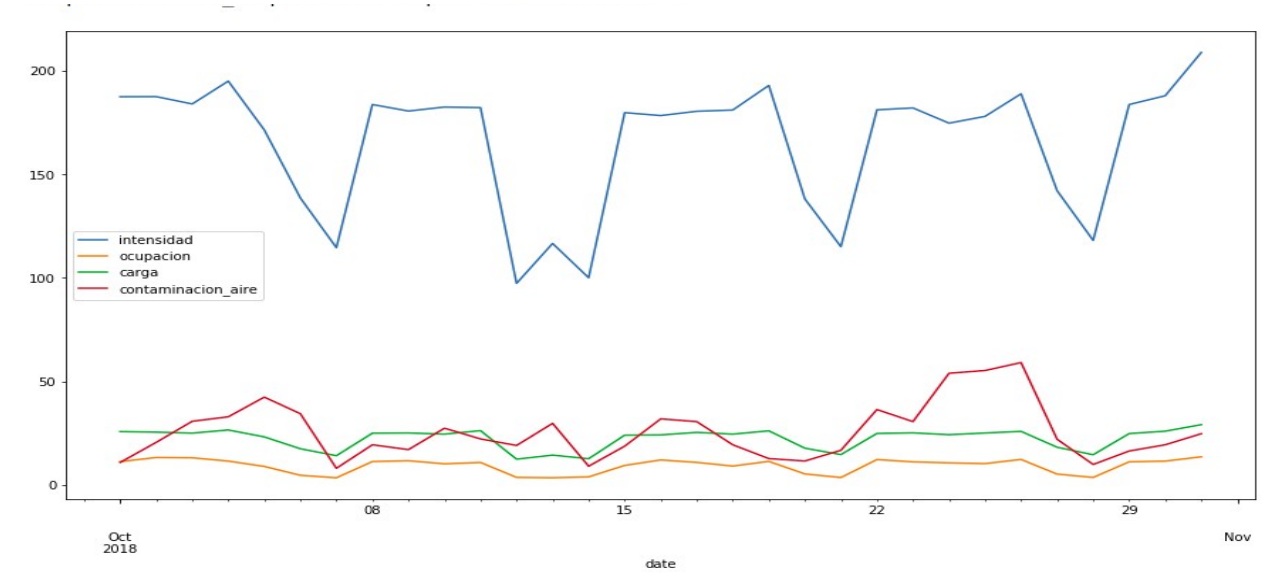
Estado del tráfico

6. 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.

7. De nuevo, tomando la ID de Moratalaz se eliminan todos los NaNs. La columna *error* tiene N en todos sus valores, lo que indica que todos los datos han sido recogidos sin errores ni sustituciones.

Comparaciones

En estas condiciones, se han representado gráficamente los resultados, sin observar gran relación entre el estado del tráfico diario y la contaminación de ese día. Su matriz de correlación lo confirmaría:



	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

Finalmente se ha realizado una estudio de la influencia del estado del tráfico de días anteriores sobre la calidad del aire de cada día. Entre ellos el que mejor resultados ofrece es la ocupación media de los 3 días anteriores, con una correlación de `0.53`