Comparativa de la calidad del aire en relación a la cantidad de tráfico en las ciudades de Barcelona y Madrid

Ciclo de vida de los datos

21 de enero de 2019

Blanco Cacharrón, Miguel Carlos
Cerezo Fernández, Elsa
Ibrain Rodríguez, Álvaro
Pascal, Razvan
Ruiz Salmón, Julia
Traspuesto Abascal, Miguel

Repositorio: https://github.com/MiguiTE/DLC-trabajoGrupo

ÍNDICE

1. Descripción general del proyecto y análisis del problema

1.1. Interés y objetivo

Hoy en día se puede encontrar una gran cantidad de noticias sobre el calentamiento global, la contaminación y sus efectos adversos en la salud y naturaleza. Por ello, este proyecto se focaliza en la observación de la calidad del aire de las dos ciudades más pobladas de España, Madrid y Barcelona.

Actualmente, la capital sigue tomando medidas relacionadas con el tráfico, ya que aseguran que es el principal culpable de la situación actual de contaminación porque está relacionado con el aumento de un peligroso contaminante, el NO_2. Por otro lado, Barcelona, aunque no iguala a Madrid en el estado de contaminación, también está tomando medidas.

Por estas razones, el presente proyecto tiene como objetivo observar si es cierta la existencia de una relación entre la cantidad de tráfico para el empeoramiento de la calidad del aire en ambas ciudades.

1.2. Cobertura temporal y geográfica

En primer lugar, para obtener un conjunto de datos con el que poder desarrollar conclusiones consistentes, la parte del proyecto dedicada a la medición de datos tendrá una duración de dos años, comenzando el 1 de enero de 2018 y finalizando el 31 de diciembre de 2019.

Por otro lado, para llevar a cabo las mediciones en cada zona, habrá dos equipos independientes situados uno en cada ciudad y que no dispondrán del mismo presupuesto para llevar a cabo la instalación de los sensores encargados de tomar los datos. El equipo ubicado en Madrid contará con un mayor presupuesto porque, al tener mayor extensión terrenal por cubrir que el de Barcelona, tendrá que instalar una mayor cantidad de sensores para cubrir proporcionalmente ambas ciudades.

Para tener un conjunto de datos lo más completo y preciso posible, la medición de los datos del tráfico se efectuará cada quince minutos en ambas ciudades y, a diferencia de este parámetro, la calidad del aire se recogerá de forma distinta en cada ciudad: diariamente en Madrid y cada hora en Barcelona. La diferencia entre la medida de cada parámetro viene dada porque se considera ineficaz realizar la toma de datos de la calidad del aire con la misma frecuencia que los datos de tráfico, ya que un cambio en el tráfico no provocará una variación inmediata en la calidad del aire. Por otro lado, la calidad del aire de Madrid se tomará con una menor frecuencia puesto que se ha considerado que ésta será suficiente para obtener resultados satisfactorios.

1.3. Requisitos técnicos

En este apartado se especifican los artefactos que serán necesarios para llevar a cabo el proyecto de forma exitosa.

En primer lugar, serán necesarios sensores para realizar las mediciones de la calidad del aire y la cantidad de tráfico. Para el primer tipo de medición se necesitarán analizadores de contaminantes (ej. Hidrocarburos, óxidos de nitrógeno, etc.) y para el segundo, sólo serán necesarios sensores que midan el tránsito.

Por otra parte, para almacenar los datos, que aproximadamente ocuparán un total de 15 GB, se necesitará un servidor de al menos 20GB de capacidad con un sistema gestor de bases de datos NoSQL y una conexión de 3-4G para mandar los datos de los sensores al servidor.

Por lo tanto, para la instalación de los elementos que se necesitarán para la realización del proyecto se estima la siguiente tabla de presupuestos:

Tipo sensor	Coste por sensor	Ciudad	Cantidad de sensores	Subtotal
Sensor que mide el tránsito	150 €	Madrid	4200	630.000 €
		Barcelona	600	90.000 €
Sensor que analiza la calidad del aire	50 €	Madrid	25	1.250 €
		Barcelona	10	500 €
			Presupuesto sensores:	721.750 €

Por otro lado, se utilizarán los servicios en la nube de Amazon para mantener en funcionamiento un servidor de 20 GB de capacidad, durante las 24 horas del día, los 365 del año y contando los dos años que dura el proyecto. Teniendo en cuenta estas especificaciones, según la estimador de coste de la página oficial de Amazon implicará un coste total de 400,70 € al año.

Por lo tanto, sin contar el coste humano, el proyecto se estima que supondrá un coste de 722.550 €.

2. Descripción de las fuentes de datos

2.1. Fuentes de datos de la ciudad de Madrid

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:

2.1.1. 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)

2.1.2. 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'

2.1.3. 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

2.1.4. 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. N: no ha habido errores ni sustituciones. E: los parámetros de calidad de algunas de las muestras integradas no son óptimos. 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.

2.2. Fuentes de datos de la ciudad de Barcelona

Los datos que se utilizan en la parte del proyecto orientada a la recolección de datos de la ciudad de Barcelona se han obtenido de diversos sensores colocados en la ciudad.

A continuación, se explican las fuentes utilizadas así como la estructura de los datos elegidos.

2.2.1. Situación de estaciones de medición de la calidad del aire

Se trata del conjunto de datos que contiene las localizaciones y datos de las estaciones de medición en Barcelona.

Fuente:

http://opendata-ajuntament.barcelona.cat/data/es/dataset/qualitat-aire-estacions-bcn

Estructura:

Nombre	Descripción
Nom_cabina	Nombre de la estación
Codi_dtes	Identificador de la estación
Zqa	Código de la zona
Codi_eoi	Código europeo de la cabina
Longitud	Longitud geográfica
Latitud	Latitud geográfica
Ubicació	Nombre de la calle
Codi_Districte	Código del distrito
Nom_Districte	Nombre del distrito
Codi_Barri	Código del barrio
Nom_Barri	Nombre del barrio
Ocupacio_sol	Tipo de suelo (Urbana, Suburbana ó Rural)
Emissions_properes	Emisiones dominantes en la zona (Tráfico ó Industrial)
Contaminant_1	Indicador contaminante NO_2
Contaminant_2	Indicador contaminante O_3
Contaminant_3	Indicador contaminante PM 10

2.2.2. Mediciones de la calidad del aire

Este conjunto de datos contiene las medidas realizadas por las estaciones en tiempo real.

Fuente:

http://opendata-ajuntament.barcelona.cat/data/es/dataset/qualitat-aire-detall-bcn

Estructura:

Nombre	Descripción
Nom_cabina	Nombre de la estación que ha tomado la medida
Qualitat_aire	Calidad del aire (Buena, Regular o Pobre)
Codi_dtes	Identificador de la estación
Zqa	Código de la zona
Codi_eoi	Código europeo de la cabina que ha tomado la medida
Longitud	Longitud geográfica
Latitud	Latitud geográfica
Hora_o3	Hora de la medición de O3 (cada hora)
Qualitat_o3	Calidad del índice O3
Valor_o3	Valor de la medida O3
Hora_no2	Hora de la medición de NO2 (cada hora)
Qualitat_no2	Calidad del índice NO2 (Buena, Regular o Pobre)
Valor_no2	Valor de la medida de NO2
Hora_pm10	Hora de la medición de PM 10 (cada hora)
Qualitat_pm10	Calidad de las partículas en suspensión (Buena, Regular o Pobre)
Valor_pm10	Valor de la medida de PM 10
Generat	Fecha y hora de cuándo se ha generado la medida
DataTime	Timestamp de la hora de creación del fichero

2.2.3. Tramos de carretera

El conjunto de datos que se puede encontrar en el siguiente enlace contiene los detalles de cada tramo de carretera de Barcelona.

Fuente:

http://opendata-ajuntament.barcelona.cat/data/ca/dataset/transit-relacio-trams/resource/036bfde0-b73e-4cb9-93db-5785f032ab68

Estructura:

Nombre	Descripción
Tram	Identificador del tramo de medición
Descripción	Nombre del tramo
Coordenadas	Coordenadas de inicio y fin de cada tramo (lat0, lon0, lat1,lon1)

2.2.4. Mediciones del estado del tráfico

Todas las medidas realizadas por los sensores de tráfico con una frecuencia de 5 minutos son regidas en el conjunto de datos al que redirige el enlace que aparece más abajo.

Fuente:

http://opendata-ajuntament.barcelona.cat/data/es/dataset/trams

Estructura:

Nombre	Descripción
IdTram	Identificador del tramo de medición
Data	Fecha de la medición
EstatActual	Estado actual (0=nada, 1=muy fluido, 2=fluido, 3=denso, 4=muy denso, 5=congestionado, 6=atasco)
EstatPrevist	Estado previsto en 15 minutos (0=nada, 1=muy fluido, 2=fluido, 3=denso, 4=muy denso, 5=congestionado, 6=atasco)

3. Data Management Plan

El plan de gestión de datos realizado en este proyecto sigue la plantilla para proyectos de la Comisión Europea (H2020) y puede encontrarse en el Anexo 1.

4. Diagrama de Gantt

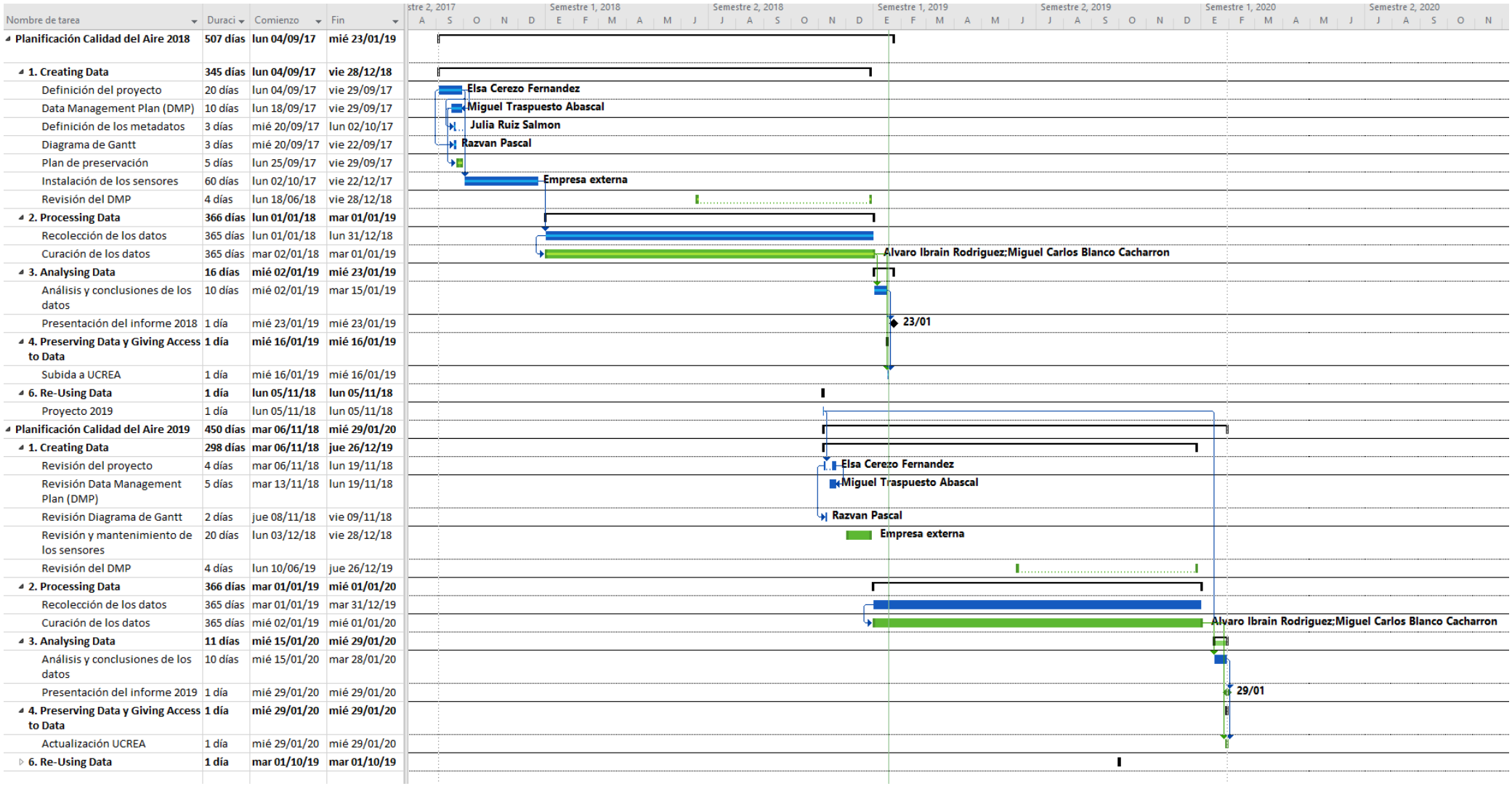

Ilustración 1. Diagrama de Gantt del proyecto.

Como se puede ver en el diagrama anterior, la representación de las tareas se ha realizado agrupándose las mismas en las diferentes etapas que contempla la aproximación UK Data Archive DLC. Además, las tareas han sido divididas en dos planificaciones: la del año 2018 y la del 2019.

A continuación, se definen de manera escueta las tareas, su duración y sus dependencias:

PLANIFICACIÓN DEL PROYECTO CALIDAD DEL AIRE: 2018

1. Etapa de creación de los datos:

1.1. Definición del proyecto [20 días]: Se define el proyecto completo sobre la Calidad del Aire en las ciudades de Madrid y Barcelona para los años 2018 y 2019.
1.2. Data Management Plan (DMP) [10 días]: Documento que forma parte del proyecto, teniendo una dependencia con el mismo de Fin a Fin (FF).

1.3. Definición de los metadatos [3 días]: Se debe incluir en el DMP, dependencias con T1.2 de Comienzo a Comienzo con un retardo de 2 días (CC+2días).

1.4. Diagrama de Gantt [3 días]: Se necesita para la definición del proyecto. Dependencias con T1.1 del tipo CC+2días.

1.5. Plan de preservación [5 días]: Se debe incluir en el DMP. Dependencias con T1.2 del tipo CC+5días.

1.6. Instalación de los sensores [60 días]: No se puede realizar si no han finalizado T1.1 y T1.2. Dependencias de Fin a Comienzo (FC).

1.7. Revisión del DMP [4 días]: Se contempla la revisión del DMP como medida de control antes de comenzar el segundo año de proyecto.

2. Etapa de procesado de datos:

2.1. Recolección de los datos [365 días]: Se recolectan los datos obtenidos de los sensores, siendo necesaria la instalación de estos previamente. Dependencia con T1.6 del tipo FC.
2.2. Curación de los datos [365 días]: La curación de los datos se realizará forma paralela a la recolección. Dependencia con T2.1 del tipo CC+1día.

3. Etapa de análisis de los datos:

3.1. Análisis y conclusiones [10 días]: Se necesita disponer de todos los datos ya curados. Dependencia con T2.2 del tipo FC.
3.2. Presentación del informe 2018 [1 día]: La presentación de los resultados no se puede realizar sin el análisis de los datos. Dependencias con T3.1 del tipo FC. En el Diagrama se representa como Hito.

4. Etapa de preservación de los datos y de acceso a los datos:

4.1. Subida a UCrea [1 día]: Se suben los datos a UCrea y se les da acceso público a los datos. Dependencias con T2.2 y T3.1. del tipo FC.

5. Etapa de reutilización de los datos:

5.1. Proyecto 2019 [1 día]: Siendo el alcance del proyecto de dos años, se usarán los resultados del proyecto como inputs del siguiente.

PLANIFICACIÓN DEL PROYECTO CALIDAD DEL AIRE: 2019

1. Etapa de creación de los datos:

1.1. Revisión del proyecto [4 días]: Se revisa la definición del proyecto.
1.2. Data Management Plan (DMP) [5 días]: Revisión del DMP.
1.3. Revisión diagrama de Gantt [2 días]: Revisar planificación del proyecto.
1.4. Revisión y mantenimiento de los sensores [60 días]: Se realiza una revisión del estado de los sensores y su mantenimiento.
1.6. Revisión del DMP [4 días]: Última revisión del DMP.

2. Etapa de procesado de datos: [Idéntica a 2018]

2.1. Recolección de los datos [365 días]: [Idéntica a 2018]
2.2. Curación de los datos [365 días**]:** [Idéntica a 2018]

3. Etapa de análisis de los datos:

3.1. Análisis y conclusiones [10 días]: Se realiza un análisis conjunto de los datos de 2018 y 2019. Dependencia con T2.2 del tipo FC (de ambos años).
3.2. Presentación del informe 2018 [1 día]: [Idéntica a 2018]

4. Etapa de preservación de los datos:

4.1. Actualización UCrea [1 día]: Se actualizan los datos de UCrea añadiendo los de 2019 y, por lo tanto, los datos añadidos pasan a tener accesibilidad pública. Dependencias con T2.2 y T3.1. del tipo FC.

5. Etapa de reutilización de los datos:

5.1. Proyecto 2019 [1 día]: Siendo el alcance del proyecto de dos años, se usarán los resultados del proyecto como inputs del siguiente.

5. Data Curation

5.1. Proceso de curación de los datos de Madrid

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:

https://github.com/MiguiTE/DLC-trabajoGrupo/blob/master/Doc/InformeCurationMadrid.pdf

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 estudio se realiza durante este mes, de forma que se obtenga 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.

5.2. Proceso de curación de los datos de Barcelona

En este apartado, se describen los pasos realizados para la curación de los datos descritos en el primer apartado del informe, que se han realizado con el objetivo de unirlos para obtener, dada una estación de medición, la información del estado del aire y del estado del tráfico cercano.

Los pasos que se muestran a continuación se pueden ver resueltos en el Jupyter Notebook dataCurationBarcelona.ipybn recogido en el repositorio del proyecto:

https://github.com/MiguiTE/DLC-trabajoGrupo/blob/master/Doc/InformeCurationBarcelona.pdf

Pasos realizados durante la curación de los datos

1. Por medio de la librería Pandas, en Python, se han cargado los archivos .csv obtenidos del portal de datos en abierto de Barcelona.
2. Se han convertido las fechas en formato String a DateTime con el fin de poder hacer comparaciones y agrupaciones por fecha.
3. Se han mapeado los valores de calidad del aire de Strings a Ints con el objetivo de realizar cálculos numéricos sobre ellos.

4. Para los tramos de carretera, se ha añadido una nueva columna, PuntoMedio, la cual contiene el punto medio de las coordenadas de cada tramo.
5. En base al punto medio y a las coordenadas de las estaciones, se ha asignado a cada tramo una estación cercana. Es decir, se ha añadido una nueva columna llamada NearStation con el identificador de la estación a la que pertenece.
6. Se han agrupado los datos de calidad por estación, día y hora, de tal manera que se puede acceder a la media de calidad de aire horaria por cada estación.
7. Sabiendo el día y el identificador de la estación, se pueden consultar los datos del tráfico con media horaria.

6. Metadatos

En este proyecto se han creado metadatos para recoger información sobre distintos elementos implicados siguiendo el formato *Dublin Core* extendido.

Los metadatos generados son:

- Metadatos sobre el proyecto
- Metadatos sobre los datos de la calidad del aire de Madrid
- Metadatos sobre los datos de la calidad del aire de Barcelona
- Metadatos sobre los datos del tráfico de Madrid
- Metadatos sobre los datos del tráfico de Barcelona

Se pueden acceder a todos ellos en la dirección https://github.com/MiguiTE/DLC-trabajoGrupo/blob/master/Metadata/MetadataDublinCore.ipynb

7. Plan de preservación

Para la preservación de los datos recolectados se utilizará el repositorio UCrea para almacenarlos, ya que es un proyecto en colaboración con la Universidad de Cantabria.

Por otro lado, tanto la seguridad como el estado de los datos dependerá de UCrea, por ello se mantendrán accesibles hasta que UCrea considere cambiar su estado.

8. Análisis de datos

Tras todo el proceso de recolección y curación de datos, se realiza la etapa de análisis en la que primero se van a comentar los resultados hallados en la observación de la ciudad de Madrid y, después, los de la ciudad de Barcelona.

- **Resultado de la observación de la ciudad de Madrid**

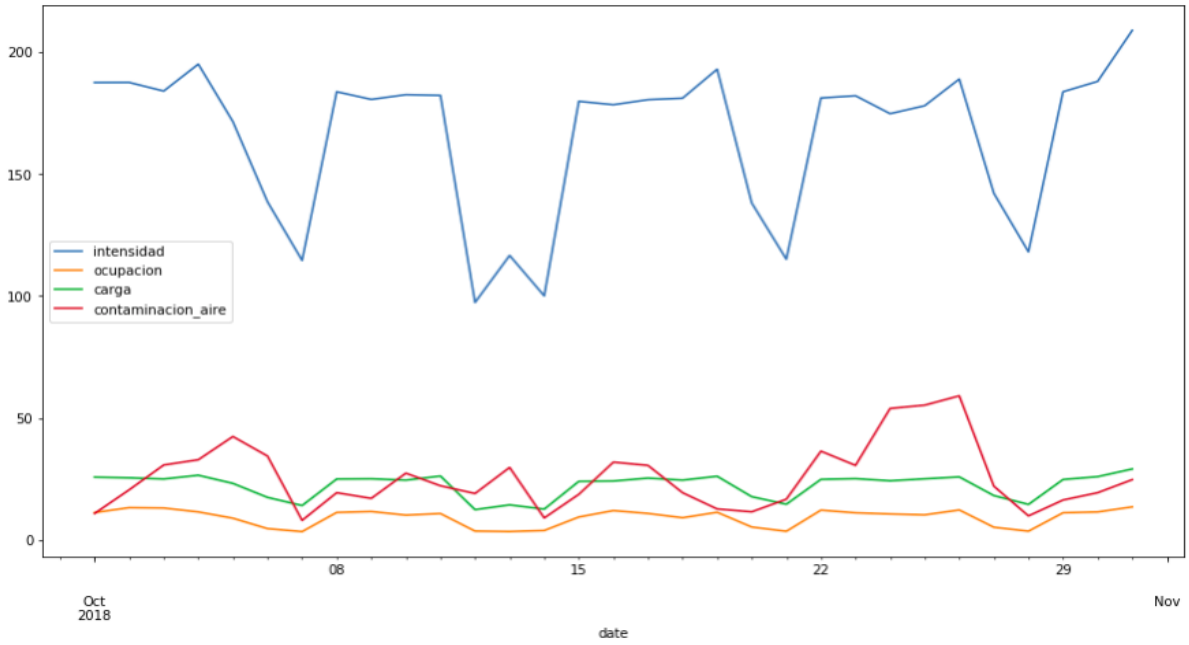

Ilustración 2. Comparación contaminación y estado del tráfico de octubre de 2018 en Madrid.

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 3. Matriz de correlación entre distintos factores del tráfico y la contaminación de Madrid según Ilustración 2.

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.

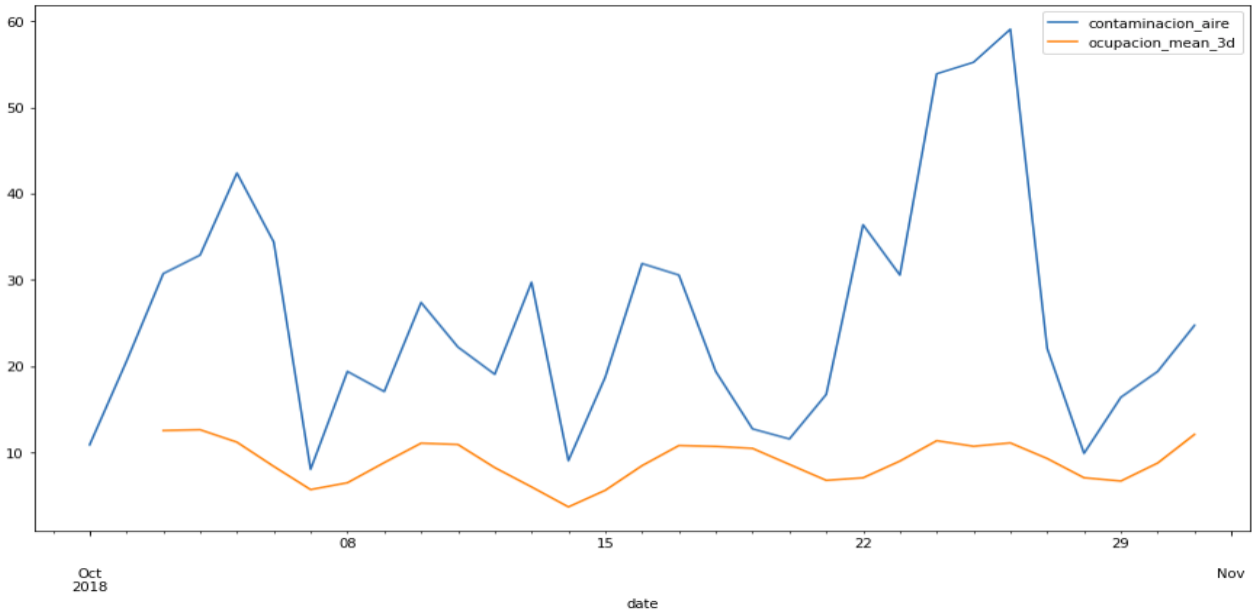

Ilustración 4. Comparación contaminación del aire con la media de la ocupación del tráfico de los tres días anteriores en octubre de 2018

- **<u>Resultado de la observación de la ciudad de Barcelona</u>**

En este caso, en el siguiente gráfico, se puede ver claramente una relación entre el tráfico y la calidad del aire. Sobre las seis de la mañana, el tráfico comienza a aumentar y unas horas más tarde, debido a la acumulación de gases y partículas contaminantes, empeora la calidad del aire.

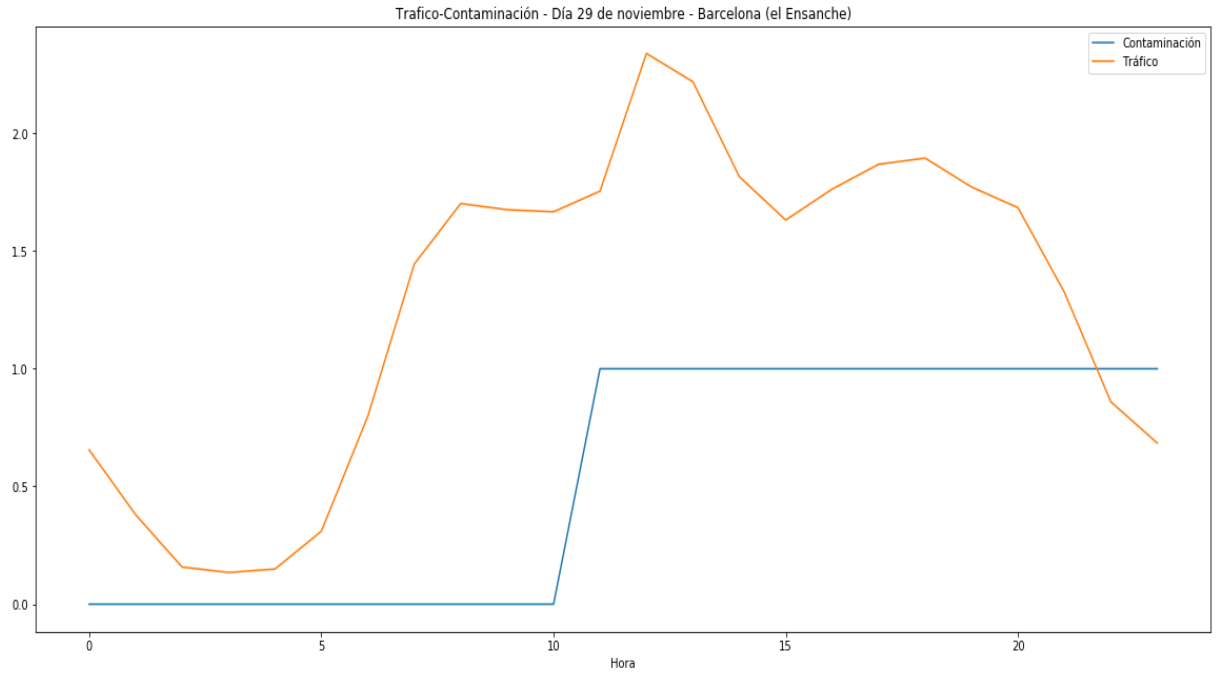

Ilustración 5. Comparación entre el tráfico y calidad del aire durante el día 29 de noviembre de 2018 en Barcelona.

ANEXO 1

Data Management Plan

(También puede encontrarse en la dirección: https://github.com/MiguiTE/DLC-trabajoGrupo/blob/master/Doc/DataManagementPlan.pdf)

1. Data Summary

¿Cuál es el propósito de la recolección/generación de datos y su relación con los objetivos del proyecto?

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Justificar la creencia de que en las grandes ciudades con un gran volumen de tráfico existe una menor calidad del aire que se respira. Para ello se realiza una comparativa de calidad de aire y volumen de tráfico en las ciudades de Madrid y Barcelona.

¿Qué tipo y formatos de datos se generarán/recolectarán?

What types and formats of data will the project generate/collect?

Se generarán datos en archivos de extensión *.csv* (*comma separated values*) compuesto de valores, generalmente en *comma flotante*. Además, habrá coordenadas geográficas para las estaciones de medición. Y un breve texto descriptivo de las mismas.

¿Se reusarán datos ya existentes? ¿Cómo?

Will you re-use any existing data and how?

No se reusarán los datos existentes, ya que son íntegramente generados en este proyecto.

¿Cuál es el origen de los datos?

What is the origin of the data?

Los datos son generados mediante sensores instalados en las ciudades de Madrid y Barcelona.

¿Cuál es el tamaño esperado de datos?

What is the expected size of the data?

A continuación, se realiza una estimación del tamaño que ocupará el conjunto de datos por medición:

- **Mediciones de la calidad del aire en Madrid**

Nombre	Descripción	Bytes	Tipo
PROVINCIA	Código de la provincia	4	Int
MUNICIPIO	Código del municipio	4	Int
ESTACIÓN	Código de la estación	4	Int
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)	3	String
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	13	String
ANO	Año medida	4	Int
MES	Mes medida	4	Int
D0i	Contaminación del día i	1	Int
V0i	Validación de la contaminación. Solo validos los registros con 'V'	1	Char

En Madrid se tomarán datos de la calidad del aire cada día y se estima que tendrá un tamaño aproximado de 38 bytes. Entonces, si se multiplica el tamaño de una toma de datos por las 25 estaciones de medición que se estiman instalar, resultan unos 950 bytes, aproximadamente, que tendrán que almacenarse en el servidor diariamente.

- **Mediciones del tráfico en Madrid**

Nombre	Descripción	Bytes	Tipo
Id	Identificador único del punto de medida	4	Int
Fecha	Fecha con formato yyyy-mm-dd hh:mm:ss	8	Date
Tipo_elem	Nombre del tipo de punto de medida: Urbano o M30	7	String
Intensidad	Intensidad del punto de medida en el periodo de 15 minutos (vehículos/hora). Un valor negativo implica la ausencia de datos.	4	Float
Ocupación	Tiempo de ocupación del punto de medida en el periodo de 15 minutos (%). Un valor negativo implica la ausencia de datos.	4	Int
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.	1	Int
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.	4	Int
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 algunas de las muestras integradas no son óptimos. S: Alguna de las muestras recibidas era totalmente errónea y no se ha integrado.	1	Char
Periodo_integración	Número de muestras recibidas y consideradas para el periodo de integración.	4	Int

En este último caso, se supone que se instalarán unos 4200 sensores que realicen la toma de medida de los datos del tráfico de Madrid. Cada toma se realizará con una frecuencia de 15 minutos y con un peso aproximado de 37 Bytes. Por lo tanto, se espera que el servidor tendrá que recibir una cantidad de 14 MB diarios.

- **Mediciones de la calidad del aire en Barcelona**

Nombre	Descripción	Bytes	Tipo
Nom_cabina	Nombre de la estación que ha tomado la medida	60	String
Qualitat_aire	Calidad del aire (Buena, Regular o Pobre)	7	String
Codi_dtes	Identificador de la estación	4	Int
Zqa	Código de la zona	4	Int
Codi_eoi	Código europeo de la cabina que ha tomado la medida	4	Int
Longitud	Longitud geográfica	4	Float
Latitud	Latitud geográfica	4	Float
Hora_o3	Hora de la medición de O3 (cada hora)	7	String
Qualitat_o3	Calidad del índice O3	7	String
Valor_o3	Valor de la medida O3	4	Float
Hora_no2	Hora de la medición de NO2 (cada hora)	7	String
Qualitat_no2	Calidad del índice NO2 (Buena, Regular o Pobre)	7	String
Valor_no2	Valor de la medida de NO2	4	Float
Hora_pm10	Hora de la medición de PM 10 (cada hora)	7	String
Qualitat_pm10	Calidad de las partículas en suspensión (Buena, Regular o Pobre)	7	String
Valor_pm10	Valor de la medida de PM 10	4	Float
Generat	Fecha y hora de cuándo se ha generado la medida	8	Datetime
DataTime	Timestamp de la hora de creación del fichero	8	Datetime

Cada toma de datos de la calidad del aire de Barcelona se realizará cada hora y se estima que tendrá un tamaño de 157 bytes. Esto resulta que se almacenarán, aproximadamente, 4 KB diarios por cada estación.

Por lo tanto, con el supuesto de colocar 10 estaciones habrá que almacenar un total de 37 KB diarios.

- **Mediciones del tráfico en Barcelona**

Nombre	Descripción	Bytes	Tipo
IdTram	Identificador del tramo de medición	4	Int
Data	Fecha de la medición	8	Datetime
EstatActual	Estado actual (0=nada, 1=muy fluido, 2=fluido, 3=denso, 4=muy denso, 5=congestionado, 6=atasco)	1	Int
EstatPrevist	Estado previsto en 15 minutos (0=nada, 1=muy fluido, 2=fluido, 3=denso, 4=muy denso, 5=congestionado, 6=atasco)	1	Int

Cada medida del tráfico de Barcelona se tomará cada 15 minutos y se estima que cada una tendrá un tamaño de 14 bytes. Por lo tanto, si además se tiene en cuenta que se estima una cantidad de 600 sensores, resulta que tendrán que poder almacenarse 0,77 MB diarios, aproximadamente.

Por lo tanto, en total se espera que el servidor debe soportar aproximadamente 15MB diarios, lo que supondrá unos 11 GB en los 2 años de duración del proyecto.

¿A quién le resultará útil ('data utility')?

To whom might it be useful ('data utility')?

- En el ámbito de la política, para contemplar nuevas leyes contra la contaminación de la zona y así colaborar contra el cambio climático.
- Para estudios científicos sobre el aire.
- Para empresas que quieran producir nuevos servicios/productos: coches menos contaminantes.

2. Fair data

- ***Making data findable, including provisions for metadata***

¿Se pueden encontrar los datos producidos y/o usados en el proyecto con metadatos? ¿Se pueden identificar y localizar mediante los estándares de identificación (¿p.e. identificadores persistentes y únicos como DOI?

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Se han escrito metadatos que describen el proyecto, así como los datos producidos. Además, al subir la documentación del proyecto junto con los datos utilizados al repositorio de UCrea, éste generará automáticamente un identificador digital, persistente y único que puede encontrarse en los metadatos del proyecto.

¿Cuáles son los estándares de nomenclatura que se siguen?

What naming conventions do you follow?

Se va a usar el estándar de metadatos *Dublin Core* extendido para la generación de los metadatos. A continuación, se listan los cambios realizados respecto a la versión original de *Dublin Core:*

En los metadatos del proyecto se han realizado lo siguientes cambios:

1. En la etiqueta *attribute* se ha añadido un atributo *xml, name,* que añade información extra de lo que contiene esta etiqueta. Por ejemplo:
   - <dc:attribute name="Campo Barcelona">O3: Ozono</dc:attribute>
   - <dc:attribute name="Campo Madrid">CONTAMINACION_AIRE: magnitud del parámetro de contaminación</dc:attribute>

   Estos atributos indican medidas que se han tomado, en el contenido de la etiqueta, y a dónde pertenecen, en el contenido del atributo.
2. En la etiqueta *source* se ha añadido un atributo *xml, name,* indicando a dónde pertenece el recurso, es decir, a Madrid o Barcelona.

En cuanto a los metadatos de los datos:

1. Tanto en *source* como en *identifier* se ha añadido el atributo *name* para indicar a qué pertenece el recurso o identificador dentro del campo.
    - <dc:source name="Tráfico">http://opendata-ajuntament.barcelona.cat/data/es/dataset/trams</dc:source>
2. Además, se ha creado el atributo *legend* para incluir la leyenda de los datos.

¿Se proporcionan palabras clave que optimizarán las posibilidades de reutilización?

Will search keywords be provided that optimize possibilities for re-use?

Sí. Se listan a continuación:

- Calidad aire Madrid
- Tráfico carretera Madrid
- Contaminación Madrid
- Medio ambiente
- Calidad aire Barcelona
- Tráfico carretera Barcelona
- Contaminación Barcelona
- Salud

¿Se ofrece un control de versiones claro?

Do you provide clear version numbers?

Todo el proceso que hayan sufrido los datos, desde su toma hasta su estudio está descrito en el repositorio de GitHub. Por lo tanto, el control de versiones vendrá dado por la herramienta Git, que proporciona dicho control.

¿Qué metadatos van a generarse? En el caso de que no exista ningún estándar de metadatos en la disciplina, por favor, describe qué tipos de metadatos son creados y cómo.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Como se ha dicho anteriormente, se crearán metadatos de los conjuntos de datos que se listan a continuación, siguiendo el estándar de *Dublin Core* extendido:

- Metadatos sobre el proyecto
- Metadatos sobre los datos de la calidad del aire de Madrid
- Metadatos sobre los datos de la calidad del aire de Barcelona
- Metadatos sobre los datos del tráfico de Madrid
- Metadatos sobre los datos del tráfico de Barcelona

- ***Making data openly accessible***

¿Qué datos producidos y/o usados en el proyecto serán publicados abiertamente? Si ciertos conjuntos de datos no se pueden compartir (o son repartidos bajo ciertas restricciones), explicar por qué, claramente separando las razones legales y contractuales provenientes de restricciones voluntarias.

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Será posible acceder a públicamente a todo el proyecto, tanto a los datos, como a la documentación. Además, en este caso los conjuntos de datos no tienen restricciones legales, por lo tanto, serán completamente accesibles.

¿Cómo se harán los datos accesibles (p.e. mediante repositorio)?

How will the data be made accessible (e.g. by deposition in a repository)?

Como el proyecto será llevado a cabo junto con la Universidad de Cantabria, los datos y el proyecto serán subidos al repositorio UCrea.

¿Qué métodos o programas informáticos son necesarios para acceder a los datos?

What methods or software tools are needed to access the data?

Se puede acceder a través un navegador a la página web de UCrea.

Por otro lado, como los archivos en los que se van a almacenar serán de formato *.csv* no hará falta ningún software para poder verlos, cualquier bloc de notas será suficiente. Sin embargo, se recomienda utilizar hojas de cálculo tipo *Excel*, *LibreOffice Calc* o *software* tipo *Pandas* de *Python, R*.

Si hay restricciones en el uso, ¿cómo se dará acceso?

If there are restrictions on use, how will access be provided?

Los datos sobre la calidad de aire y el volumen de tráfico serán accesibles públicamente y fuera de restricciones legales. Por lo que serán totalmente accesibles.

- ***Making data interoperable***

¿Son los datos producidos en el proyecto interoperables, esto es, se permite el intercambio y reciclado de datos entre investigadores, instituciones, organizaciones, países etc. (i.e. agregar formatos estándar, tanto como es posible complementar con *software* accesible (*open)*, y en particular facilitar combinar con diferentes conjuntos de datos de diferentes orígenes)?

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and facilitating re-combinations with different datasets from different origins)?

Sí, ya que los datos serán guardados en formato *.csv* y, además, accesibles públicamente por internet.

Además, se proporcionan metadatos de los datos para así facilitar lo máximo posible la combinación con otros datos de distintos orígenes. Aunque el formato de metadatos, *Dublin Core* extendido, no es estándar, las modificaciones que se ha hecho son mínimas y se describen para facilitar su comprensión.

¿Qué vocabularios de datos y metadatos, estándares o metodologías se seguirán para hacer los datos interoperables?

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Metadatos: Dublin Core extendido.

Metodologías: Se incluye código escrito en *Python,* con el paquete *Pandas*.

- ***Increase data re-use (through claryfying licenses)***

¿Cuál será la licencia de los datos para permitir el mayor reciclado posible?

How will the data be licensed to permit the widest re-use possible?

La licencia de los datos será *Creative Commons 4.0.* Resumiendo, esta licencia permite:

- **Compartir** — copiar y redistribuir el material en cualquier medio o formato
- **Adaptar** — volver a mezclar, transformar y crear a partir del material para cualquier finalidad, incluso comercial.

Siempre y cuando se cumpla:

- **Reconocimiento** — Debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios. Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.

No hay restricciones adicionales — No puede aplicar términos legales o medidas tecnológicas que legalmente restrinjan realizar aquello que la licencia permite.

¿Cuándo serán accesibles los datos para reciclado? Sin embargo, hace que la publicación se retrase, problemas con patentes, especificar por qué y cuánto tiempo se hará, con la intención de que los datos deberán ser accesibles tan pronto como sea posible.

When will the data be made available for re-use? If an-embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Como la publicación del proyecto y los datos se realizará en el repositorio de UCrea, una vez se publiquen se podrá acceder a ellos. La parte del proyecto del año 2018 tiene como fecha estimada para su publicación a partir del 16 de enero de 2019. Sin embargo, la relacionada con el año 2019 no estará disponible hasta el 29 de enero de 2020 según la estimación realizada.

Por otro lado, el estudio será gestionado por dicho repositorio y, por lo tanto, no se sabe cuánto tardará en solucionarse en este momento.

¿Cuánto tiempo se pretende que los datos puedan ser reciclados?

How long is it intended that the data remains re-usable?

El tiempo que el repositorio UCrea mantenga accesible los datos.

¿Se describen los procesos que aseguran la calidad de los datos?

Are data quality assurance processes described?

Sí. Como se ha dicho anteriormente, se incluyen los *scripts* de todos los procesos que se han hecho a la hora de trabajar con los datos.

3. Allocation of resources

¿Cuáles son los costes de hacer los datos de vuestro proyecto *FAIR*?

What are the costs for making data FAIR in your project?

Los datos ya cumplen los estándares FAIR:

- **Encontrable (findability):** Se proporcionan identificadores persistentes y únicos en los metadatos.
- **Accesible (accessibility):** Es accesible públicamente mediante el repositorio UCrea.
- **Interoperable (interoperability):** Se proporcionan los datos junto con metadatos siguiendo el estándar *Dublin Core* extendido explicado para facilitar la interoperabilidad.
- **Reciclable (reusability):** Los datos son accesibles públicamente para que cualquiera pueda acceder a ellos y usarlos en futuros proyectos.

4. Data security

¿Cómo se prevee la seguridad de los datos (incluyendo recuperación de datos, así como seguridad de almacenamiento y transferencia de datos personales)? ¿Están los datos almacenados de forma segura en repositorios certificados para la preservación y curación a largo plazo?

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)? Is the data safely stored in certified repositories for long term preservation and curation?

La seguridad de los datos y el estudio dependerá totalmente de la de la Universidad de Cantabria. No compete a este proyecto hacerse cargo de su seguridad.

5. Ethical aspects

¿Existen problemas legales o éticos que puedan causar problemas a la hora de compartir los datos? Estos pueden ser discutidos en el contexto de la revisión ética. Si es relevante, incluye referencias al capítulo de ética en la DoA

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Los datos no tienen ninguna información sensible que pueda suponer problemas éticos o legales.

¿En los cuestionarios existe información para el consentimiento de preservación a largo plazo, así como compartir los datos personales?

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

No procede. No hay datos personales.