

Universidades de Burgos, León y
Valladolid

Máster universitario

Inteligencia de Negocio y Big Data en Entornos Seguros



**TFM del Máster Inteligencia de Negocio y Big
Data en Entornos Seguros**

**Estudio analítico sobre la calidad del
aire en la ciudad de Madrid**

Presentado por Adrián Aguado García

—
30 de agosto de 2019

Tutor Académico: Carlos E. Vivaracho Pascual

Tutor Empresa: Fernando Cuenca Cabezas

Universidades de Burgos, León y Valladolid



Máster universitario en Inteligencia de Negocio y Big Data en Entornos Seguros

Carlos E. Vivaracho Pascual, profesor de la Universidad de Valladolid, del departamento de Informática, área de Ciencias de la Computación y la Inteligencia Artificial.

Expone:

Que el alumno D. Adrián Aguado García, con DNI 78759314T, ha realizado el Trabajo final de Máster en Inteligencia de Negocio y Big Data en Entornos Seguros titulado "Estudio analítico sobre la calidad del aire en la ciudad de Madrid".

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 30 de agosto de 2019

Vº. Bº. del Tutor:

D. nombre tutor
Carlos E. Vivaracho Pascual

Vº. Bº. del co-tutor:

D. nombre co-tutor
Fernando Cuenca Cabezas

Resumen

La contaminación ambiental supone, y cada vez más, un problema para la salud humana en todo el mundo. Los núcleos urbanos grandes resultan ser una de las áreas con mayor contaminación. Más concretamente diversos gases, como el dióxido de nitrógeno NO₂, están haciendo que las administraciones se planteen diversas medidas para su reducción. En el presente documento se realiza un estudio basado en un modelo analítico para los datos de contaminación de la ciudad de Madrid.

Descriptores

Big Data, Visualización de datos, Python, R, contaminación ambiental...

Abstract

Environmental pollution is a problem for human health worldwide. Large urban centres are one of the most polluted areas. More specifically, various gases, such as nitrogen dioxide NO₂, are causing administrations to consider various measures for their reduction. This document is a study based on an analytical model for pollution data from the city of Madrid.

Keywords

Big Data, Data Visualization, Python, R, pollution . . .

Índice general

Índice general	III
Índice de figuras	v
Índice de tablas	vii
Memoria	1
Introducción	2
Objetivos del proyecto	4
2.1. Objetivos	4
2.2. Objetivos personales	4
Conceptos teóricos	6
3.1. Introducción	6
3.2. Big data	6
3.3. Análisis científico	10
3.4. Análisis legislativo (En España)	26
3.5. Análisis parte técnica	28
Técnicas y herramientas	33
4.1. Metodologías	33
4.2. Herramientas	35
4.3. Herramientas de visualización de datos	36
4.4. Tecnologías	38
4.5. Documentación	39
4.6. Otras herramientas	40

Aspectos relevantes del desarrollo del proyecto	43
5.1. Elección del proyecto	43
5.2. Formación	43
5.3. Metodologías/Estrategias aplicadas	44
5.4. Desarrollo del modelo analítico	44
5.5. Desarrollo visualización	44
5.6. Desarrollo web	45
5.7. Documentación	45
5.8. Dificultades encontradas	45
5.9. Agradecimientos	45
Trabajos relacionados	47
6.1. Trabajos útiles	47
Entorno experimental	49
7.1. Esquema	49
7.2. Fuente datos origen	50
7.3. Exploración	55
7.4. Transformación de datos	58
7.5. Visualización de datos	60
Resultados	65
8.1. Preámbulo	65
8.2. Resultados generales obtenidos	66
8.3. Preguntas analíticas	66
Conclusiones y Líneas de trabajo futuras	68
9.1. Conclusiones proyecto	68
9.2. Conclusiones personales	68
9.3. Líneas de trabajo futuras	68
Apéndices	69
Bibliografía	70
Apéndice A Plan de Proyecto Software	73
A.1. Introducción	73
A.2. Planificación temporal	73
A.3. Viabilidad legal	82
A.4. Links Importantes	84

Índice de figuras

3.1.	Arq. tradicional vs Arq. <i>Big Data</i> . Fuente: https://www.slideshare.net/bd4s/big-data-introduccion	7
3.2.	Emisiones de CO ₂ mundiales. Fuente: Infografía por @countcarbon	11
3.3.	Tabla ICA. Fuente: https://www.elespanol.com/omicrono	13
3.4.	Mapa estaciones calidad del aire España. Fuente: Metadatos gobierno España	16
3.5.	Calidad del aire por contaminante. Fuente: Informe Calidad Aire 2018 [18]	16
3.6.	Mapa España superación límite legal durante 2018 con Madrid en rojo. Fuente: Ecologistas en acción [7]	20
3.7.	Localización estaciones Madrid. Fuente: Elaboración propia mediante cartodb y los datos de madrid [9]	21
3.8.	Perímetro Madrid Central. Fuente: Ayto Madrid	25
3.9.	Respuesta ayuntamiento de Madrid. Fuente: Consulta realizada Ayto Madrid	28
3.10.	Tipos de datos a descargar (datos diarios) Fuente: Datos Madrid	29
3.11.	<i>What makes a good visualization?</i> by David McCandless. Fuente: InformationIsBeautiful.net	31
4.12.	Resumen metodología <i>Scrum</i> . Fuente: manifesto.co.uk	34
4.13.	<i>Board</i> de tareas empleado durante el proyecto. Fuente: Elaboración propia	35
4.14.	Pantalla principal Zotero. Fuente: elaboración propia	40
4.15.	Detalle vista principal. Fuente: elaboración propia	42
7.16.	Detalle diagrama desarrollo (Partes 1 y 2). Fuente: Elaboración propia	49
7.17.	Detalle diagrama desarrollo (Partes 3 y 4). Fuente: Elaboración propia	50
7.18.	Detalle <i>vega-voyager</i> . Fuente: Elaboración propia	56
7.19.	Detalle <i>Datawrapper</i> . Fuente: Elaboración propia	56
7.20.	Datos cargados en <i>tinybird</i> . Fuente: Elaboración propia	59
7.21.	Detalle <i>tinybird</i> . Fuente: Elaboración propia	59
7.22.	Detalle <i>tinybird</i> . Fuente: Elaboración propia	60
7.23.	Detalle <i>tinybird</i> . Fuente: Elaboración propia	60
7.24.	Detalle NO ₂ 2019 (julio). Fuente: Elaboración propia	61
7.25.	Detalle NO ₂ 2019 (12 julio). Fuente: Elaboración propia	62
7.26.	Detalle <i>page</i> en <i>PowerBi</i> carga. Fuente: Elaboración propia	63

7.27. Detalle <i>page</i> en <i>PowerBi</i> carga. Fuente: Elaboración propia.	63
7.28. Ejemplo gráfico por tamaño. Fuente: Elaboración propia.	64
7.29. Ejemplo gráfico nube de palabras.	64
8.30. Diagrama desarrollo. Fuente: Elaboración propia.	65
A.1. Detalle <i>Product Backlog</i>	75
A.2. Detalle <i>Scrum board</i>	76
A.3. Detalle Sprint 1.	78
A.4. Detalle Sprint 2.	80
A.5. Detalle Sprint 3.	81
A.6. Gráfico licencias Open Source in GitHub. Fuente: https://cartograf.net . . .	83
A.7. Licencia Creative Commons.	84

Índice de tablas

3.1.	Tipos de medidores y sus contaminantes principales [2]	14
3.2.	Tabla valores límite [18]	18
3.3.	Tabla valores objetivo [18]	19
3.4.	Tabla información Madrid. Fuente: INE [18]	19
7.5.	Tabla intérprete valores horarios.	52
7.6.	Tabla intérprete valores diarios.	53
7.7.	Tabla intérprete estaciones de control.	54
7.8.	Tabla intérprete contaminantes.	55
7.9.	Tabla tipo datos valores horarios.	57
7.10.	Tabla tipo datos valores diarios.	57
A.1.	Tabla resumen-licencias.	82

Memoria

Introducción

La contaminación ambiental se ha incrementado, y sigue haciéndolo a un ritmo vertiginoso, de manera preocupante en los últimos años. Constituye uno de los problemas más serios a los que se enfrenta el ser humano. Hoy día ya no es una cuestión localizada en algunos lugares sino que el viento se ha encargado de convertirlo en un problema global. Los gases provenientes de automóviles, camiones, procesos industriales, sistemas de calefacción e incluso hasta el humo de los cigarrillos de miles de fumadores se juntan para contaminar el aire que consumimos a diario.

Respirar aire limpio y sin riesgos para la salud debería ser un derecho de toda persona. Está demostrado que la contaminación atmosférica causa graves daños a la salud y al medio ambiente. Los niveles actuales de contaminación atmosférica provocan la muerte de entre cuatro y cinco millones de personas por exposición directa al aire contaminado en todo el mundo [29]. Si hablamos de España alrededor de 16.000 muertes prematuras en España [19].

Pero la contaminación es un tema realmente amplio, dado que la mayor parte de contaminantes son expulsados por procesos industriales y automovilísticos; y además éstos se concentran principalmente en grandes urbes, es precisamente ahí donde debemos realizar nuestros mayores esfuerzos. Casi la mitad de la población mundial vive actualmente en ciudades, y para el año 2050 se prevé que aumente a un 75 % [3]. Nuestro reto ahora es el intentar cambiar la manera en la que vivimos en aquellos puntos de mayor concentración de contaminación para evitar seguir contaminando de la manera que lo hacemos. Globalmente ya se están tomando medidas, diversos países proponen medidas para tratar de frenar o reducir la contaminación.

En el caso concreto de la ciudad de Madrid, se trata de una de las dos ciudades españolas, junto con Barcelona, que está obligada a cumplir los niveles máximos de NO₂ impuestos como medidas para reducir la contaminación dictados por la unión europea.

En las páginas siguientes vamos a realizar un estudio mediante el cuál

analizaremos la calidad de aire de la ciudad de Madrid gracias al portal de datos abiertos y basándonos en los niveles de calidad del aire que recomienda la unión europea como normales.

Objetivos del proyecto

En este apartado se van a detallar los diferentes objetivos que se buscaban con la realización de este trabajo fin de máster.

Se parte de la hipótesis de que es posible obtener datos abiertos que permitan caracterizar los niveles de contaminación de la ciudad de Madrid y que a partir de estos datos se puede crear un modelo para obtener, analizar y visualizar los datos de la calidad del aire.

2.1. Objetivos

Comenzaremos con los objetivos generales del proyecto.

- Analizar, usando los datos abiertos que proporciona el Ayuntamiento de Madrid, la calidad del aire de la ciudad, poniéndola en relación con los objetivos planteados por la Unión Europea
- Recopilación, descripción, exploración, visualización
- Preparación de los datos (Selección, limpieza, transformación, preprocesado)
- Visualización de datos que ayude a la respuesta de diversas preguntas analíticas

2.2. Objetivos personales

Los objetivos personales que he perseguido durante todo el desarrollo han sido los siguientes:

- Realizar un trabajo fin de máster más enfocado a la investigación. Durante mi trabajo fin de grado realicé una aplicación web y móvil. Podemos decir que aquel trabajo era algo más puramente técnico por lo que cuando vi la posibilidad de algo que combinara la parte de exploración,

al final ha sido necesario involucrarse e indagar en la parte científica del trabajo, con la parte de tecnología, todo lo aprendido en el máster, lo escogí sin dudar. Además no descarto el realizar un doctorado en un futuro y creo que me va venir bien comenzar a tener esta visión más investigadora.

- Realización de un proyecto que involucre datos reales
- Mejorar mis conocimientos de Python, R, visualización de datos y estadística. Dado que mi trabajo actual engloba el desarrollo web, he escogido esos lenguajes y estas particularidades por que deseo enfocar mi futuro profesional hacia este campo.
- Profundizar en ciencia de datos
- Profundizar en visualización de datos

Conceptos teóricos

3.1. Introducción

Esta sección pretende servir como orientación antes de adentrarse en el trabajo en sí mismo. Se trata de un conglomerado de consideraciones de diferente índole que es necesario conocer para comprender este trabajo fin de máster.

3.2. Big data

La información siempre ha sido, y será, el tesoro más valioso con el que puede contar una empresa. En la era de la sociedad de la información y en la que, al día, se generan 2,5 quintillones ¹ de datos, resulta más que evidente que se trata de la herramienta empresarial y social más importante de la historia.

Empresas de cualquier tamaño pueden y deben usar esta tecnología para cubrir prácticamente cualquier necesidad que tengan. Se trata de un nuevo prisma que modifica la visión de un negocio desde cualquier perspectiva. Sin embargo, podemos afirmar que se trata de un área de reciente implantación, que avanza a una velocidad vertiginosa, y que pocas empresas emplean con resultados altamente satisfactorios. Por todo ello no existen definiciones estrictamente formales del término. Una aproximación a una interpretación del término podría ser la siguiente:

Big Data refers to a data set that is so large and/or complex that it cannot be perceived, acquired, managed, and processed by traditional Information Technology (IT) and software/hardware tools within a tolerable time [23]

¹Fuente IBM, Noticia ABC. <https://www.abc.es/tecnologia/redes/20140423/abci-trillones-byte-informacion-cada-201404222207.html>

En otras palabras, *Big data* es un concepto en auge que se puede describir como una gran masa de datos cuyo gran tamaño hace complejo su análisis con las técnicas más habituales. El problema actual reside en que, y vamos a más, los grandes volúmenes de datos han superado con creces a la capacidad de procesamiento de un simple *host*.

Arquitecturas

La arquitectura de un proyecto *Big Data* está compuesta generalmente por varias capas, es posible encontrar fuentes que hablan de cuatro y también de cinco, aunque quizás es verdad que depende del proyecto al que se aplican. Particularmente creo que está mejor enfocado con cuatro capas que serían las siguientes: *recolección de datos*, *almacenamiento*, *procesamiento de datos* y *visualización*. Esta arquitectura “por capas” no es nueva, sino que ya es algo generalizado en las soluciones de *Business Intelligence* que existen hoy en día. Sin embargo, debido a las nuevas necesidades cada uno de estos pasos ha ido adaptándose y aportando nuevas tecnologías a la vez que abriendo nuevas oportunidades.

En la figura 3.1 podemos ver de manera resumida la diferencia entre una arquitectura clásica y una arquitectura *Big Data*.

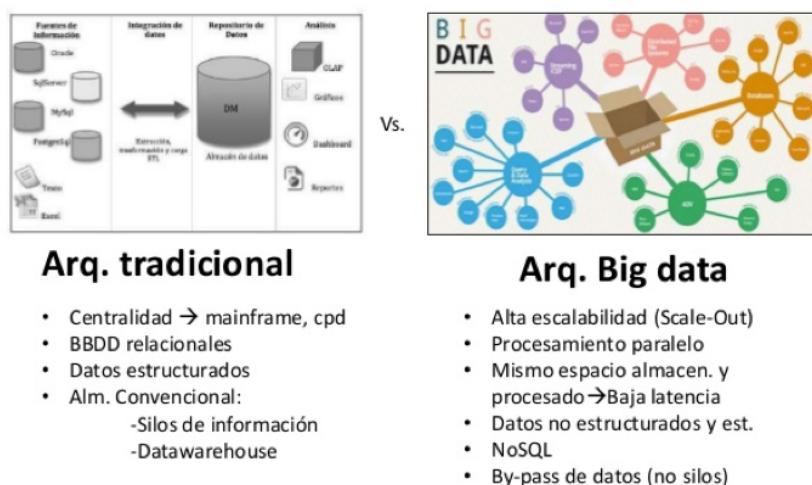


Figura 3.1: Arq. tradicional vs Arq. *Big Data*. Fuente: <https://www.slideshare.net/bd4s/big-data-introduccion>

Smart cities

Aunque es práctica invisible para nosotros, cantidades ingentes de información se almacenan y circulan alrededor de nuestro mundo cada día generado mayormente por las empresas, administraciones o los propios ciudadanos en

una explosión de datos a la espera de ser procesados y aprovechados con fines diversos. Desde hace unos años, la transformación digital que estamos sufriendo todos supone un cambio de paradigma en nuestras vidas. Existen multitud de actividades, por no decir casi todas, donde la influencia del análisis de datos nos proporciona predicciones realmente interesantes. Si este tipo de actividades se realizan dentro de las ciudades, es decir, se aplica a la habitabilidad y al diseño de las mismas, entonces estaremos hablando de lo que los expertos llaman *Smart* [20].

Las ciudades ofrecen distintos servicios a los ciudadanos por lo cual necesitan recopilar y almacenar una gran cantidad de datos. Otros organismos y administraciones públicas, financiados por la ciudadanía, generan a su vez información geográfica, cartográfica, meteorológica, médica etc. Los partidarios del *Open Data*, la filosofía de datos abiertos, defienden que esta información debería ser accesible y reutilizable para el público en general, sin exigencia de ningún permiso específico. El por qué parece bastante lógico ya que consideran que restringir su acceso va contra el bien común debido a que se trata de información que pertenece a la sociedad, o ha sido financiada por ella. Por lo tanto, intentan promover que ese tipo de datos estén accesibles para todo el mundo que así lo desee.

Hablamos de *Smart Cities* cuando las ciudades tratan de innovar para mejorar la calidad de vida de sus habitantes. Se trata de un término cada vez más común y hace referencia a sistemas de infraestructura, conectividad, *IoT* (Internet of things), y, por supuesto, datos abiertos que conviven interconectados entre sí en un núcleo de población. Si queremos conocer algunas de estas ciudades nada mejor que visitar el informe anual de la escuela de negocios IESE, llamado [IESE Cities in Motion](#). Un gran ejemplo de ese tipo de ciudades es Nueva york.

***Big Data* y ciudades**

Hasta ahora nunca habíamos obtenido y recogido tal cantidad de datos en las ciudades, y lo que es quizás más importante, tenemos la capacidad para procesarlos y entenderlos. En las ciudades, aunque quizás jamás nos hayamos parado a pensar, es usual la producción de conjuntos de datos enormes. Además esto no resulta específico de hoy en día, sino desde hace tiempo se realizan censos, encuestas, entrevistas, baremos... El ser humano y sus acciones generamos datos pero las administraciones también. De la misma manera las empresas generan y almacenan cantidad de datos provenientes de sus operaciones internas. Sin embargo estamos hablando de que todos estos datos suelen ser muy específicos por lo que resultan ser complejos a la hora de su análisis. Dependiendo de la manera de cómo se generan estos datos, es decir del origen que tienen dentro de las ciudades, podríamos tener datos *dirigidos* (for-

mas tradicionales de recopilado como registros médicos), *automatizado* (datos recogidos mediante sensores) o *voluntario* (datos extraídos de redes sociales).

Como ya hemos nombrado, la población humana presenta una clara tendencia a concentrarse en los núcleos urbanos². Esta cantidad de personas hace que sea necesario el actuar en base a estos datos para intentar mejorar la calidad de vida de las personas que viven en esas ciudades.

Hoy en día las administraciones de grandes ciudades, en España tenemos ejemplos como Barcelona, Sevilla o Madrid, apuestan por un análisis en tiempo real de diferentes métricas relativas a las ciudades. Uno de los ejemplos más útiles es el del transporte público en tiempo real proporcionando a los ciudadanos implicaciones directas ya que pueden observar el estado de su próximo autobús o metro en una parada determinada y a una hora concreta. A continuación nombraremos otros ejemplos. En la lista se incluyen tanto sistemas que se están desarrollando actualmente como propuestas que se pueden realizar, o se están realizando pero no he encontrado ejemplos, en diversas ciudades gracias al análisis de los datos que en ellas se generan.

- **Sistemas de transporte:** los problemas de movilidad, con el caso de transporte público o la congestión de tráfico, es posible paliarlos por medio de sistemas inteligentes de regulación de tráfico. Entre los medios utilizados para recabar datos de los agentes involucrados en el tráfico en las ciudades y de las infraestructuras viarias, destacan las videocámaras y diferentes tipos de sensores.
- **Seguridad:** sin duda es una de las ventajas más directas ya que través de cámaras de videovigilancia o sensores es posible controlar a los delincuentes de una manera más eficaz. Londres por ejemplo cuenta con más de 40.000 cámaras de vigilancia aunque hay quien critica que quizás no sea la metodología más adecuada [8]
- **Gestión de residuos:** en diversas ciudades ya se utilizan sistemas inteligentes para optimizar la gestión de residuos tanto para la parte de clasificación como para la parte de recogida. [4]
- **Sanciones:** una de las aplicaciones más rentables para las administraciones en sin duda la del control del cumplimiento de las normas de circulación. Algunos ejemplos que actualmente se pueden realizar de manera automática son invadir carriles reservados, pisar una línea continua, incumplir una señal de stop o saltarse un semáforo en rojo. Cámaras con reconocimiento facial y de matrículas permiten identificar el vehículo y también al conductor infractor.

²Fuente: Banco Mundial. <https://datos.bancomundial.org/indicador/sp.urb.totl.in.zs>

- **Sistemas medioambientes:** todo lo referente al medio ambiente se está utilizando desde hace años y se va a continuar utilizando para tratar de optimizarlo al máximo. Algunos de estos sistemas utilizan tecnología ubicua [11], capaz de suministrar información sobre diversas magnitudes de interés.
- **Energía eléctrica:** las empresas privadas de energía ya llevan años recopilando datos gracias a los contadores.
- **Sistema de abastecimiento de agua:** el derroche de agua es un problema en muchas ciudades y con sistemas basados en datos extraídos a partir de sensores de presión sería posible controlar la cantidad y calidad de agua en tiempo real.
- **Eficacia de servicios públicos:** se me ocurre también la eficacia en actuaciones de cuerpos de seguridad, ambulancias o bomberos a través de la correlación de toda la información procedente de diversos sistemas y accesible a través de herramientas en tiempo real.

En el caso de este trabajo nos centraremos en la parte de sistemas medioambientales o emisiones nocivas. Mediante sensores ubicados por toda la geografía de una ciudad se recogen de manera periódica diversos datos sobre diferentes contaminantes que resultan ser nocivos para la salud. Asimismo también existen medidores de temperatura, precipitaciones u otro tipo de agentes medioambientales.

3.3. Análisis científico

Se define la contaminación atmosférica como “*la presencia en la atmósfera de materias, sustancias o formas de energía que impliquen molestia grave, riesgo o daño para la seguridad o la salud de las personas, el medio ambiente y demás bienes de cualquier naturaleza*”.

La atmósfera es un bien común indispensable para la vida. Dada su condición esencial para la vida humana y por los daños que su contaminación o deterioro puede causar, la calidad del aire y la protección de la atmósfera ha sido, desde hace décadas, una prioridad política ambiental a lo largo de todo el mundo. Podemos afirmar que la contaminación atmosférica es consecuencia directa de las emisiones al aire de los gases y material particulado derivados de la actividad humana (social y económica) y de fuentes naturales [18].

Si hablamos del mundo, China lidera el ranking de países emisores de dióxido de carbono (CO_2), seguido de EE. UU. y de los estados miembros de la UE. El potente gas de efecto invernadero es una de las sustancias que más contribuyen al calentamiento global y al cambio climático, pero no es la única.

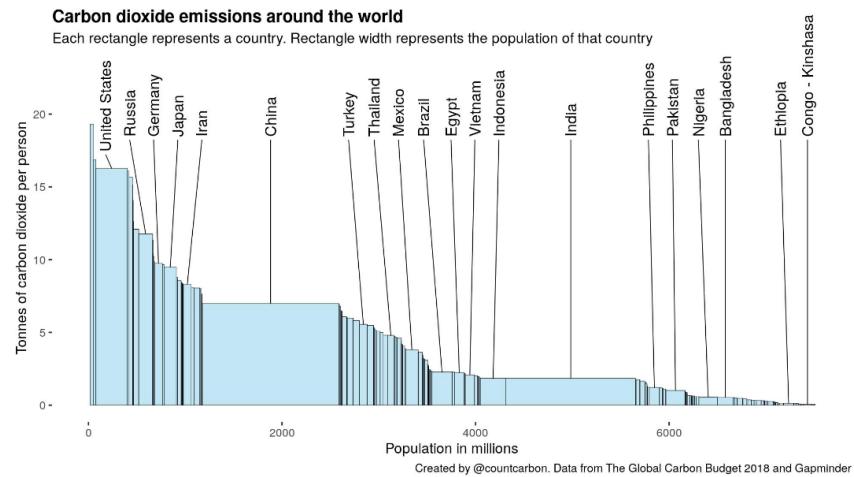


Figura 3.2: Emisiones de CO₂ mundiales. Fuente: Infografía por [@countcarbon](#)

La figura 3.2 muestra qué países contaminan y cuánto: en vertical están las toneladas de CO₂ por persona y en horizontal la población de cada país, de modo que el rectángulo simboliza las emisiones totales. China se lleva la palma, aunque sus emisiones son la mitad por persona que las de Estados Unidos. Hasta Alemania o Japón están peor (pero tienen menos personas y emisiones totales).

En cuanto a europa, las últimas estimaciones globales de la AEMA ([Agencia Europea de Medio Ambiente](#)) y la OMS ([Organización Mundial de la Salud](#)) sobre la repercusión sanitaria de la contaminación atmosférica son muy preocupantes. Elevan en el año 2015 hasta medio millón las muertes prematutras en los países europeos por la mala calidad del aire, 422.000 por exposición a partículas inferiores a 2,5 micras de diámetro (PM_{2,5}), 79.000 por exposición a dióxido de nitrógeno (NO₂) y 17.700 por exposición a ozono troposférico. En España, las víctimas de la contaminación serían ya más de 30.000 al año, 27.900 por partículas PM_{2,5}, 8.900 por NO₂ y 1.800 por ozono, lo que supone duplicar los 16.000 fallecimientos prematuros anuales que se estimaban hace apenas una década [7].

El coste económico de la mortalidad prematura y de la pérdida de días de trabajo por la contaminación del aire ambiente y en el interior de las viviendas ha sido cuantificado por el Banco Mundial en 38.000 millones de euros en 2013, equivalentes al 3,5 por ciento del Producto Interior Bruto (PIB) español, sin considerar los daños provocados a los cultivos, los ecosistemas naturales u otros bienes de cualquier naturaleza [7].

¿Cómo se mide la calidad del aire?

La calidad del aire viene determinada principalmente por la distribución geográfica de las fuentes de emisión de contaminantes y las cantidades de contaminantes que se emiten. Pero ¿qué es un contaminante? Según la directiva europea *2008/50/CE* (CE, 2008), un contaminante es “toda sustancia presente en el aire ambiente que pueda tener efectos nocivos para la salud humana y el medio ambiente en su conjunto”.

Cualquier método que permita medir, calcular, predecir o estimar las emisiones, los niveles o los efectos de la contaminación atmosférica resulta válido para llevar a cabo la evaluación de la calidad del aire. Sin embargo dependiendo de los tipos de contaminantes en los que nos queremos enfocar existen metodologías diferentes. Esto es debido a qué no todas las metodologías aportan la misma precisión, se tienen en cuenta diversos factores.

A finales de 2017, la Agencia Europea de Medio Ambiente ([AEMA](#)) y la Comisión Europea pusieron en marcha un nuevo Índice europeo de Calidad del Aire ([ICA](#)), que permite a los usuarios comprobar su calidad en las más de 2.000 estaciones de medición repartidas por toda Europa. Este índice, el cuál se basa en la figura 3.3, proporciona información actualizada sobre la calidad del aire en los 33 países miembros de la AEMA, incluye perfiles nacionales ya que las administraciones públicas locales tienen que adaptar sus medidas para considerar factores como demografía, infraestructuras de transporte, etc. Estas mediciones vigilan los estándares de calidad del aire y controlan los niveles de ozono (O_3), de dióxido de nitrógeno (NO_2), de dióxido de carbono (CO_2), de dióxido de azufre (SO_2)... y toda la contaminación generada por partículas que pueden representar serios riesgos para la salud.

La agencia de medioambiente [EPA](#) es la responsable del índice en EE. UU. mientras que el Centro Nacional para la Monitorización del Medio ambiente en China (CNMMC) es el organismo responsable de compilar, analizar, agregar y publicar los datos de los diferentes indicadores del aire en el país asiático.

ICA	COLOR	CLASIFICACIÓN	O ₃ 8h ppm	O ₃ 1h ppm	PM ₁₀ 24h µg/m ³	PM _{2,5} 24h µg/m ³	CO 8h ppm	SO ₂ 24h ppm	NO ₂ 1h ppm
0 - 50	Verde	Buena	0.000 0.059	-	0 54	0 12	0 4.4	0 0.035	0 0.053
51 - 100	Amarillo	Moderada	0.060 0.075	-	55 154	12.1 35.4	4.5 9.4	0.036 0.075	0.054 0.100
101 - 150	Naranja	Dañina a la salud para grupos sensibles	0.076 0.095	0.125 0.164	155 254	35.5 55.4	9.5 12.4	0.076 0.185	0.101 0.360
151 - 200	Rojo	Dañina a la salud	0.096 0.115	0.165 0.204	255 354	55.5 150.4	12.5 15.4	0.186 0.304	0.361 0.649
201 - 300	Púrpura	Muy Dañina a la salud	0.116 0.374	0.205 0.404	355 424	150.5 250.4	15.5 30.4	0.305 0.604	0.650 1.249
301 - 400	Marrón	Peligrosa	-	0.405 0.504	425 504	250.5 350.4	30.5 40.4	0.605 0.804	1.250 1.649
401 - 500	Marrón	Peligrosa	-	0.505 0.604	505 604	350.5 500.4	40.5 50.4	0.805 1.004	1.650 2.049

Figura 3.3: Tabla ICA. Fuente: <https://www.elespanol.com/omicrono>

Esta contaminación suele darse cuando las condiciones atmosféricas son favorablemente agradables (cuando hay sol y el cielo está despejado). En estos casos, el suelo se calienta durante el día y se enfriá durante la noche, de modo que el aire se estanca y no se regenera. El diseño de la ciudad también puede hacer aumentar o disminuir la polución, haciendo que el aire se estanke o se regenere en mayor o menor medida.

El método técnico para saber la cantidad de contaminación que hay es mediante estaciones meteorológicas, también conocidas como estaciones de seguimiento de contaminación o estaciones remotas de medición de la calidad del aire. Estas estaciones miden la concentración de distintos agentes contaminantes en el aire. En la tabla 3.3 se resumen algunos de los gases y los medidores que son necesarios para en análisis de los mismos.

Contaminante a analizar	Descripción medidores
Dióxido de azufre (SO_2)	Este sistema se basa en la fluorescencia producida por las moléculas del dióxido de azufre.
Monóxido de carbono (CO)	Este sistema se basa en la radiación infrarroja que absorben las moléculas de monóxido de carbono.
Ozono (O_3)	Basado en la radiación ultravioleta.
Óxido de nitrógeno (NO)	Se basan en la energía que libera la unión de óxido de nitrógeno con ozono.
Partículas en suspensión ($\text{PM}_{2,5}$)	Este tipo de dispositivos analizan partículas en suspensión cuyo diámetro sea mayor a $2,5 \mu\text{m}$.
Hidrocarburos (CH)	Estos dispositivos detectan las partículas que libera la combustión del hidrógeno.

Tabla 3.1: Tipos de medidores y sus contaminantes principales [2]

Agentes contaminantes principales

Hablando en sentido general, los contaminantes no siempre son producidos por el hombre sino también por algunos fenómenos naturales, por ejemplo las erupciones volcánicas. En este breve apartado se va a incluir una lista con los agentes contaminantes principales [2]. Según su procedencia los podemos agrupar como sigue a continuación.

- Primarios: proceden de las fuentes de emisión
 - *Gaseosos*
 - Dióxido de azufre (SO_2)
 - Monóxido de carbono (CO)
 - Óxidos de nitrógeno (NO_x)
 - Hidrocarburos (HC)
 - Dióxido de carbono (CO_2)
 - *No gaseosos*
 - Partículas: de índole natural, en su mayoría, su procedencia es muy variada
 - *Otros*
 - Restos de combustión de fuel, gas-oil o alquitranes
 - Erupciones volcánicas
 - Incendios
 - Intrusiones de material particulado
 - Incineraciones no depuradas de basuras

- ◊ Metales pesados: plomo, cadmio, mercurio... etc.
- Secundarios: originados en la atmósfera como consecuencia de reacciones químicas que transforman los contaminantes primarios
 - Ozono (O_3)
 - Trióxido de azufre (SO_3)
 - Ácido sulfúrico (H_2SO_4)
 - Dióxido de nitrógeno (NO_2)
 - Ácido nítrico (HNO_3)

La calidad del aire en España

Todo este apartado esta basado en los diferentes informes que se realizan de manera anual por el *Ministerio de Agricultura, Alimentación y Medio Ambiente* [18]. Las siguientes líneas, que describen el marco teórico de la situación de España en cuanto a calidad de aire, se apoyan en el mencionado informe y también en el informe anual que elabora la plataforma ecologistas en acción [7].

En nuestro país se divide el territorio en zonas o aglomeraciones, en función de la densidad de población, y se evalúa la calidad del aire para los contaminantes dióxido de azufre (SO_2), dióxido de nitrógeno y óxidos de nitrógeno (NO_2 , NOX), partículas (PM_{10} y $PM_{2,5}$), plomo (Pb), benceno (C_6H_6), monóxido de carbono (CO), arsénico (As), cadmio (Cd), níquel (Ni), benzo(a)pireno (B(a)P) y ozono (O_3), que son los contaminantes con valores legislados para protección de la salud. Existen diferentes estaciones de control para la medición de estos gases repartidas por toda la península tal y como se puede ver en la figura 3.4.

Para garantizar que se abarca la totalidad de la superficie nacional, las comunidades autónomas son las encargadas de dividir su territorio en zonas de calidad del aire homogéneas para la gestión y la evaluación (mediante mediciones, modelización u otras técnicas). Para ello se determinan unos métodos y criterios comunes de evaluación. También hay que cumplir con el requisito imprescindible de informar a la población y a las organizaciones interesadas. Por lo tanto cada comunidad es la que se encarga de las mediciones particulares de cada territorio. Después, al terminar el año el gobierno se encarga de realizar el ya mencionado informe [18]. El resultado de esta evaluación anual se presenta en un cuestionario técnico para su envío a la Comisión Europea y en otros informes más claros y comprensibles dirigidos a la población.



Figura 3.4: Mapa estaciones calidad del aire España. Fuente: [Metadatos gobierno España](#)

Por motivos obvios se va a escoger como referencia el último de los informes, que es el correspondiente al periodo anual del año anterior al que nos encontramos. Los contaminantes más problemáticos en el Estado español **durante 2018** han sido las partículas en suspensión (PM_{10} y $PM_{2,5}$), el dióxido de nitrógeno (NO_2), el ozono troposférico (O_3) y el dióxido de azufre (SO_2). Se puede ver de manera más detallada en la figura 3.5.

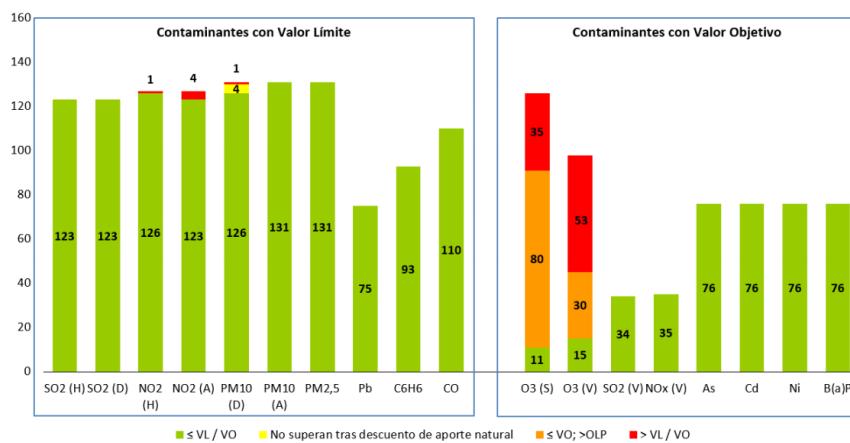


Figura 3.5: Calidad del aire por contaminante. Fuente: Informe Calidad Aire 2018 [18]

Como conclusión general de ambos informes, y en referente a la salud de

la población, se destacan dos perspectivas:

1. La población que respiró aire contaminado en el Español, según los valores límite y objetivo establecidos para los contaminantes principales citados por la *Directiva 2008/50/CE* y el Real Decreto 102/2011, alcanzó los 14,9 millones de personas, es decir un 31,8 % de toda la población. En otras palabras, **uno de cada tres españoles respiró un aire que incumple los estándares legales vigentes**. Esta situación supone no obstante un descenso de 2,6 millones de afectados respecto a 2017, y la cifra más baja desde el año 2011.
2. Si se tienen en cuenta los valores recomendados por la OMS, más estrictos que los límites legales (y más acordes con una adecuada protección de la salud), **la población que respiró aire contaminado se incrementa hasta los 45,2 millones de personas**. Es decir, un 96,8 % de la población. En otras palabras, la práctica totalidad de los españoles respiró un aire con niveles de contaminación superiores a los recomendados por la OMS. Esta situación supone un modesto descenso de 0,6 millones de afectados respecto a 2017, y se mantiene por encima de la incidencia en la década, salvo el año 2015.

Metodologías de evaluación

Independientemente de la comunidad autónoma o territorio la evaluación de la calidad del aire debe efectuarse con un enfoque común basado en criterios de evaluación también comunes. Dicha evaluación debe tener en cuenta el tamaño de las poblaciones y los ecosistemas expuestos a la contaminación atmosférica, lo que lleva a clasificar el territorio nacional en zonas o aglomeraciones en función de la densidad de población [16].

- Las zonas son porciones de territorio delimitadas por la Administración competente en cada caso utilizada para evaluación y gestión de la calidad del aire.
- Las aglomeraciones se definen como conurbaciones de población superiores a 250.000 habitantes o bien, cuando la población sea igual o inferior a 250.000 habitantes, con una densidad de población por km² que determine la Administración competente y justifique que se evalúe y controle la calidad del aire ambiente.

En las zonas y aglomeraciones así definidas se evalúa la calidad del aire para los contaminantes dióxido de azufre (SO₂), dióxido de nitrógeno y óxidos de nitrógeno (NO₂, NOx), partículas (PM₁₀ y PM_{2,5}), plomo (Pb), benceno (C₆H₆), monóxido de carbono (CO), arsénico (As), cadmio (Cd), níquel (Ni),

benzo(a)pireno (B(a)P) y ozono (O_3). La unidad utilizada en la medida de todos estos contaminantes es $\mu\text{g}/\text{m}^3$ (microgramos por metro cúbico).

Dicha evaluación se efectúa considerando diversos objetivos de calidad del aire. Se distingue entre:

- Valor límite: Objetivo para la protección de la salud, definidos para SO_2 , NO_2 , partículas PM_{10} y $\text{PM}_{2,5}$, plomo, benceno y CO.
- Valor objetivo (objetivos a largo plazo): Objetivos para la protección de la salud, definidos para partículas $\text{PM}_{2,5}$, arsénico (As), cadmio (Cd), níquel (Ni), B(a)P y ozono (O_3).
- Nivel crítico: Objetivos para la protección de la vegetación, definidos para SO_2 y NO_x .

Se entiende por valor límite aquel fijado basándose en conocimientos científicos, con el fin de evitar, prevenir o reducir los efectos nocivos para la salud humana, para el medio ambiente en su conjunto y demás bienes de cualquier naturaleza que debe alcanzarse en un período determinado y no superarse una vez alcanzado

Si nos centramos en el temas únicamente referidos a la salud existen unos **valores límite** objetivo que se recogen en la tabla 3.2.

Contaminante	Período de promedio	Valor límite
SO_2	Horario	250 $\mu\text{g} / \text{m}^3$, (máx. 24 sup. al año)
	Diario	50 $\mu\text{g}/\text{m}^3$ (máx. 3 sup. al año)
NO_2	Horario	200 $\mu\text{g}/\text{m}^3$ (máx. 18 sup. al año)
	Diario	40 $\mu\text{g}/\text{m}^3$
PM_{10}	Diario	50 $\mu\text{g}/\text{m}^3$ (máx. 35 sup. al año)
	Anual	40 $\mu\text{g}/\text{m}^3$
Pb	Anual	0,5 $\mu\text{g}/\text{m}^3$
C_6H_6	Anual	5 $\mu\text{g}/\text{m}^3$
CO	Horario	330 $\mu\text{g}/\text{m}^3$
$\text{PM}_{2,5}$	Anual	25 $\mu\text{g}/\text{m}^3$

Tabla 3.2: Tabla valores límite [18]

Si nos centramos en el temas únicamente referidos a la salud existen unos **valores objetivo** objetivo que se recogen en la tabla 3.3.

Contaminante a analizar	Descripción	Valor Objetivo
PM _{2,5}	Anual	25 µg/m ³
As	Anual	6 ng/m ³
Cd	Anual	5 ng/m ³
Ni	Anual	20 ng/m ³
B(a)P	Anual	1 ng/m ³

Tabla 3.3: Tabla valores objetivo [18]

La calidad del aire en Madrid

Ahora hablamos tan solo de la ciudad de Madrid, que posee las características descritas en la tabla 3.3.

Características		Madrid	España
Población	(Habs.)	3.354.745	46.722.980
	(%)	7,18 %	100 %
Superficie	(km ²)	7.424	505.990
	(%)	1,47 %	100 %

Tabla 3.4: Tabla información Madrid. Fuente: INE [18]

Como conclusiones de los informes del pasado año se destaca principalmente lo siguiente:

- En el año 2018 **se ha superado el valor límite anual de NO₂**, así como el valor objetivo de O₃ tanto para la protección de la salud como de la vegetación. Las causas de la superación del NO₂ se atribuyen principalmente al tráfico de vehículos de combustión ya que se trata de ubicaciones muy influenciadas por vías principales de tráfico. En la figura 3.6
- En el ámbito de esta red no se supera el valor límite horario de NO₂, pero sí se produce, sin embargo, una **superación del valor límite anual de NO₂**, concretamente en la zona ES1308 “Corredor del Henares”, como consecuencia de los niveles alcanzados en la estación ES1869A “Coslada”, de tipo urbana de tráfico (41 µg/m³ de media anual).



Figura 3.6: Mapa España superación límite legal durante 2018 con Madrid en rojo. Fuente: Ecologistas en acción [7]

Además de los valores emitidos en los informes anuales por el gobierno y otras organizaciones existen también papers, como por ejemplo [26] [22], que estudian la creciente contaminación en la ciudad de Madrid.

Madrid ciudad posee un sistema de vigilancia que dispone de una red formada por 24 estaciones que pueden clasificarse en tres categorías en cuanto al tipo de ambiente en el que se ubican: 9 estaciones de tráfico (situadas próximas a las vías), 12 estaciones de fondo urbano (más alejadas del tráfico, generalmente en parques) y 3 estaciones suburbanas (situadas fuera del núcleo urbano consolidado). En la figura 3.7 es posible observar su ubicación.

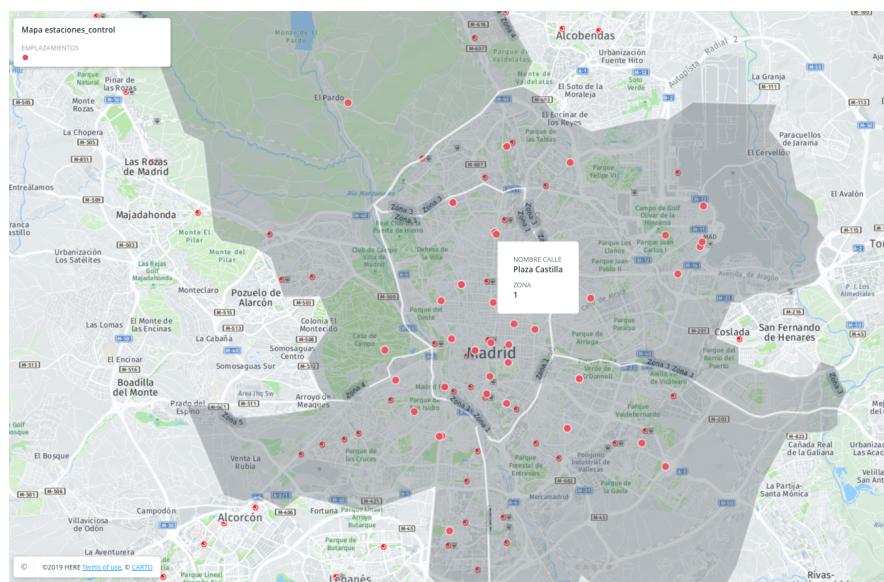


Figura 3.7: Localización estaciones Madrid. Fuente: Elaboración propia mediante [cartodb](#) y los datos de madrid [9]

Así, para el caso de los contaminantes que se analizan, tenemos que el NO₂ se mide en las 24 estaciones, las partículas PM₁₀ en 12 de ellas, las partículas PM_{2,5} en 6 y el O₃ se registra en 14 estaciones. Por otro lado, el Ayuntamiento ha establecido una zonificación de la ciudad de Madrid orientada a la gestión de situaciones de altos niveles de contaminación, como los picos de contaminación por NO₂, que ponen en marcha la aplicación del protocolo aprobado por el Ayuntamiento de Madrid para hacer frente a dichas situaciones.

Como medidas del protocolo de actuación cabe destacar que el dióxido de nitrógeno NO₂ **nunca deberá sobrepasar los 200 µg/m³**, de hacerlo se entraría en estado de ‘preaviso’.

La legislación europea, a la cual se ajusta la ciudad de Madrid, establece también un valor límite horario de NO₂, para proteger a la población de exposiciones a altos niveles de este contaminante, aunque sea por cortos períodos de tiempo (denominados “pico de contaminación”). El valor límite horario para el NO₂ está establecido en 200 µg/m³, límite que no debería rebasarse más de 18 horas al año en ninguna estación de la ciudad.

Madrid, por ser una de las ciudades más susceptibles de sufrir contaminación, tiene varios protocolos de actuación ante distintos niveles de contaminación del aire. Todos ellos se atienden al índice ICA (ver figura 3.3). Mediante este índice se indica si el aire es apto o no y qué grado de contaminación tiene.

Según el mencionado índice, si el valor está entre 0 y 50, las **condiciones**

del aire son buenas. Si se encuentra entre 51 y 100, son **regulares**. A partir de 101, y hasta el valor 150, el nivel de contaminación es **dañino** para la salud de algunos grupos (niños y ancianos, entre otros). Desde los valores 151 hasta 200, **el aire es contaminante para cualquiera**. A partir de 201, los niveles son **muy dañinos e incluso peligrosos**.

En cuanto a las situaciones de aviso existen varias:

- Preaviso: cuando dos estaciones cualesquiera detectan un nivel superior a 180 microgramos/m³ durante dos horas consecutivas.
- Aviso: cuando dos estaciones cualesquiera detectan un nivel superior a 200 microgramos/m³ durante dos horas consecutivas
- Alerta: cuando tres estaciones cualesquiera detecten un nivel superior a 400 microgramos/m³ durante tres horas consecutivas.

Durante años anteriores se han desarrollado diferentes protocolos que van en función del tipo de situaciones de aviso y de la legislación que el gobierno vigente acometa o apruebe en cada legislatura. Una vez se haya superado o se prevea superar alguno de los niveles citados en las situaciones de aviso, y si la previsión meteorológica es desfavorable, se considerará iniciado un episodio de contaminación. En ese contexto se dan varios escenarios posibles con diferentes actuaciones.

- ESCENARIO 1

* 1 día con superación del nivel de preaviso.

Actuaciones:

- Medidas informativas y de recomendación.
- Medidas de promoción del transporte público.
- Reducción de la velocidad a 70 km/h en la M-30 y accesos.

- ESCENARIO 2

* 2 días consecutivos con superación del nivel de preaviso o
1 día con superación del nivel de aviso.

Actuaciones:

- Todas las medidas del escenario 1.

- Prohibición de la circulación en el interior de la M-30 y por la M-30 a los vehículos a motor, incluidos ciclomotores, que no tengan la clasificación ambiental de “CERO EMISIONES”, “ECO”, “C” o “B” en el Registro de Vehículos de la Dirección General de Tráfico.
- Prohibición del estacionamiento en las plazas y horario del Servicio de Estacionamiento Regulado (SER) a los vehículos a motor que no tengan la clasificación ambiental de “CERO EMISIONES” o “ECO” en el Registro de Vehículos de la Dirección General de Tráfico.

■ ESCENARIO 3

- * **3 días consecutivos con superación del nivel de preaviso o 2 días consecutivos con superación del nivel de aviso.**

Actuaciones:

- Todas las medidas del escenario 1.
- Prohibición del estacionamiento en las plazas y horario del Servicio de Estacionamiento Regulado (SER) a los vehículos a motor que no tengan la clasificación ambiental de “CERO EMISIONES” o “ECO” en el Registro de Vehículos de la Dirección General de Tráfico.
- Se recomienda la no circulación de taxis libres, excepto Eurotaxis y vehículos que tengan la clasificación ambiental de “CERO EMISIONES” o “ECO” en el Registro de Vehículos de la Dirección General de Tráfico en todo el término municipal. Estos vehículos podrán estacionar en las plazas del SER, además de en sus paradas habituales a la espera de viajeros, en los términos que se establezcan en la Ordenanza de Movilidad Sostenible.

■ ESCENARIO 4

- * **4 días consecutivos con superación del nivel de aviso.**

Actuaciones:

- Todas las medidas del escenario 1.
- Prohibición del estacionamiento en las plazas y horario del Servicio de Estacionamiento Regulado (SER) a los vehículos a motor que no tengan la clasificación ambiental de “CERO EMISIONES” o “ECO” en el Registro de Vehículos de la Dirección General de Tráfico.

- Se recomienda la no circulación de taxis libres, excepto Eurotaxis y vehículos que tengan la clasificación ambiental de “CERO EMISIONES” o “ECO” en el Registro de Vehículos de la Dirección General de Tráfico en todo el término municipal. Estos vehículos podrán estacionar en las plazas del SER, además de en sus paradas habituales a la espera de viajeros, en los términos que se establezcan en la Ordenanza de Movilidad Sostenible.
- Prohibición de la circulación en el interior de la M-30 y por la M-30 a los vehículos a motor, incluidos ciclomotores, que no tengan la clasificación ambiental de “CERO EMISIONES”, “ECO” o “C” en el Registro de Vehículos de la Dirección General de Tráfico.

Como vemos son medidas acumulativas que dependen de la gravedad de la situación. Este tipo de medidas no solo permite hacer descender el nivel de contaminación durante los días que están activas, sin que también incentiva la compra de vehículos híbridos o vehículos totalmente eléctricos.

Debido a los diversos toques de atención de la Unión Europea Madrid cuenta con un protocolo de actuación claro y eficaz del que carecen otras ciudades españolas.

Low Emission Zone (LEZ): Madrid Central

También denominadas ZUAP (Zonas Urbanas de Atmósfera Protegida [13]) o ZBE (Zona de bajas emisiones), son aquellos espacios dentro de una ciudad que tienen vetada la entrada a los vehículos más contaminantes al espacio delimitado. Algunos de los países que implementan, desde hace varios años, este tipo de medidas son Noruega, Francia, Holanda o Gran Bretaña.

Madrid Central es una zona de emisiones que entró en vigor el 30 de noviembre del 2018. Esta medida está incluida dentro del Plan A de Calidad del Aire y Cambio Climático. El objetivo principal es el de reducir los gases nocivos, bajando así un 40 % las emisiones en el distrito central de la ciudad y un 20 % los desplazamientos dentro de la zona.

La zona afectada está conformada por la mayoría de calles del centro de la ciudad, entre las más relevantes incluidas se pueden destacar las siguientes: Alberto Aguilera, Glorieta de Ruíz Jiménez, Carranza, Glorieta de Bilbao, Sagasta, Plaza de Alonso Martínez, Génova, Plaza de Colón, Paseo de Recoletos, Plaza de Cibeles, Paseo del Prado, Plaza de Cánovas del Castillo, Paseo del Prado, Plaza del Emperador Carlos V, Ronda de Atocha, Ronda de Valencia, Glorieta de Embajadores, Ronda de Toledo, Glorieta de la Puerta de Toledo, Ronda de Segovia, Cuesta de la Vega, Calle Mayor, Calle Bailén, Plaza de España (lateral continuación de la Cuesta de San Vicente), Calle Princesa y Calle Serrano Jover. Podemos ver la zona delimitada en la figura 3.8.



Figura 3.8: Perímetro Madrid Central. Fuente: [Ayto Madrid](#)

La calidad del aire y la salud

El 95 % de la población mundial vive en áreas que no cumplen las pautas de un aire sano, según el informe State of Global Air [10], del Health Effects Institute. Las ciudades, donde viven más de la mitad de los casi 7.500 millones de habitantes del planeta, son el caldo de cultivo de la contaminación, un importante factor de mortalidad, al que solo superan la hipertensión, la dieta no saludable y el tabaco.

El aire contaminado puede generar problemas de pulmón o del corazón, además de problemas disfuncionales cardio respiratorios e incluso la muerte, pero esto sería únicamente ante una exposición muy prolongada, o antes varias exposiciones periódicas prolongadas. De hecho, la OMS ha advertido que representa un importante riesgo medioambiental para la salud y que provoca cada año unas tres millones de defunciones prematuras, de las cuales medio millón corresponderían a Europa [1].

Según la OMS, la contaminación del aire es actualmente uno de los mayores riesgos sanitarios mundiales, comparable a los riesgos relacionados con el tabaco [1]. Algunos ejemplos serían los siguientes.

- Limitar las vías respiratorias.
- Agravar o incluso generar enfermedades respiratorias.
- Dañar partes profundas de los pulmones, aun después de que desaparecen ciertos síntomas como tos o dolor de garganta.
- Riesgo mayor de padecer enfermedades cardiovasculares.
- Algunos de los efectos de la exposición a niveles altos de monóxido de carbono pueden disminuir los reflejos y causar confusión y somnolencia.

3.4. Análisis legislativo (En España)

Por supuesto toda esta maraña de terminología científica está regida por diversas leyes, tanto a nivel nacional como también Europeo. La presencia en la atmósfera de sustancias contaminantes, que pueden ser gases, partículas y/o aerosoles es la que determina en última instancia la calidad del aire. En España, la protección de la atmósfera y de la calidad del aire pasa por la prevención, vigilancia y reducción de los efectos nocivos de dichas sustancias contaminantes sobre la salud y el medio ambiente en su conjunto, en todo el territorio nacional. Para ello, la normativa vigente en materia de calidad del aire establece unos objetivos de calidad del aire, o niveles (concentraciones) de contaminantes en la atmósfera que no deben sobrepasarse.

España comunica anualmente información sobre calidad del aire a la Comisión Europea en cumplimiento de diferentes directivas . En la siguiente lista se realiza un recopilatorio de las más importantes a nivel europeo y nacional. En el informe anual de 2018 [18], apartado 2.1 se añade información detallada sobre el marco legislativo y cada una de las leyes que incumben a nuestro país en materia de gases nocivos.

- Marco legislativo Europeo

- *Directiva 2008/50/CE del Parlamento Europeo y del Consejo, de 21 de mayo de 2008, relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa.*
- *Directiva 2004/107/CE del Parlamento Europeo y del Consejo, de 15 de diciembre de 2004, relativa al arsénico, el cadmio, el mercurio, el níquel y los hidrocarburos aromáticos policíclicos en el aire ambiente.*

- *Directiva 2015/1480/UE, de la Comisión, de 28 de agosto de 2015, por la que se modifican varios anexos de las Directivas 2004/107/CE y 2008/50/CE del Parlamento Europeo y del Consejo en los que se establecen las normas relativas a los métodos de referencia, la validación de datos y la ubicación de los puntos de muestreo para la evaluación de la calidad del aire ambiente.*
- *Decisión de ejecución de la Comisión 2011/850/UE, de 12 de diciembre de 2011, por la que se establecen disposiciones para las Directivas 2004/107/CE y 2008/50/CE del Parlamento Europeo y del Consejo en relación con el intercambio recíproco de información y la notificación sobre la calidad del aire ambiente.*

■ Marco legislativo Nacional

- *Ley 34/2007, de 15 de noviembre, de calidad del aire y protección de la atmósfera.*
- *Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire.*
- *Orden TEC/351/2019, de 18 de marzo, por la que se aprueba el Índice Nacional de Calidad del Aire.*

A modo de resumen, en la actualidad, los textos legales más relevantes para la calidad del aire en España son: la *Directiva europea 2008/50*; la *Ley 34/2007*, de Calidad del Aire y Protección de la Atmósfera; y el *R.D. 102/2011* relativo a la mejora de la calidad del aire.

Si nos atenemos únicamente a Madrid, que es la ciudad objeto de estudio en el presente trabajo, a parte de seguir todas estas leyes nacionales y europeas poseen un protocolo de contaminación propio para actuar en caso de sobrepasar los niveles establecidos en dichas leyes. No es frecuente que cambie pero si es interesante preguntar, debido a los cambios de gobierno, cuál es el que se encuentra en vigor. Se ha consultado con la administración actual para recibir esa información se obtuvo una amable respuesta (figura 3.9). El protocolo de actuación se puede consultar de manera online [24] por cualquier ciudadano.

Madrid, a 12 de julio de 2019

Estimado señor AGUADO:

En primer lugar agradecemos que utilice el Sistema de Sugerencias y Reclamaciones del Ayuntamiento de Madrid.

En relación con la consulta que nos remite, quiero indicarle que el protocolo creado en 2015 ha sido modificado, siendo el vigente el "protocolo de actuación para episodios de contaminación por dióxido de nitrógeno en la Ciudad de Madrid" de 10 de diciembre de 2018, que podrá encontrar en la página web de esta corporación. Le adjunto el enlace:

<http://www.mambiente.munimadrid.es/opencms/opencms/calaire/Episodios/ProtocoloNO2.html>

Agradeciendo su interés le saludo atentamente,

José Amador Fernández Viejo
Director General de Sostenibilidad y Control Ambiental

Figura 3.9: Respuesta ayundamiento de Madrid. Fuente: Consulta realizada [Ayto Madrid](#)

Para ampliar la información sobre normativa se puede consultar el apartado *Acerca de Datos Abiertos / Normativa* en la [página web](#) de datos abiertos del ayuntamiento de Madrid que es el mismo que aplica a toda España.

Para terminar, y para conocer por qué me he decantado finalmente por unos contaminantes y no otros a la hora de realizar el estudio, en el *Real Decreto 102/2011* relativo a la mejora de la calidad del aire, establece umbrales de alerta para tres contaminantes, dióxido de nitrógeno (NO_2), dióxido de azufre (SO_2) y ozono (O_3). Define además el umbral de alerta como el nivel a partir del cual una exposición de breve duración supone un riesgo para la salud humana, que afecta al conjunto de la población y que requiere la adopción de medidas inmediatas. El valor del umbral de alerta para el dióxido de nitrógeno está establecido en **400 $\mu\text{g}/\text{m}^3$** durante tres horas consecutivas en lugares representativos de la calidad del aire, en un área de al menos 100 km² o en una zona o aglomeración entera, si ésta última superficie es menor.

3.5. Análisis parte técnica

En esta parte vamos a realizar una inmersión a la parte teórica de la parte más técnica del proyecto. Servirá para conocer un poco más a fondo la parte tecnológica con la que hemos realizado todo el trabajo.

Fuente de datos

Los datos provienen de [datos Madrid](#) que es el portal de datos abiertos del ayuntamiento de Madrid. La verdad que la página está realmente bien, hay *datasets* que contienen información relevante de diversa índole, se actualizan de manera frecuente y las opciones de descarga son múltiples (ver figura 3.10) lo que permite versatilidad a la hora de realizar proyectos.



Figura 3.10: Tipos de datos a descargar (datos diarios) Fuente: [Datos Madrid](#)

En este caso concreto se ha escogido el formato *.csv*. Un csv (*comma-separated values*) es un archivo de texto que almacena los datos en forma de columnas, separadas generalmente por coma y donde las filas se distinguen por saltos de línea. Este tipo de archivos podría decirse que es el lenguaje internacional de trabajo con datos, aunque por supuesto no es el único [5]. De hecho tiene un estándar [28] por el que se rige, si bien es cierto que cada persona lo interpreta un poco como conviene para cada proyecto.

En el caso concreto del estudio se han descargados los datos, en su mayoría, en formato *.csv*. Si bien es cierto que en versiones de hace más de cuatro años los formatos se limitaban solo a texto plano (*.txt*) y se han evaluado con lo que había disponible.

Modelo analítico

Alrededor de un 80 % del análisis de un proyecto o negocio se invierte en los procesos iniciales, es decir, en la elección de un buen modelo que permitir extraer, transformar y cargar los datos de una manera óptima.

Para ello lo más importante que tenemos que tener en cuenta antes de abordar un proceso que involucre grandes masas de datos a analizar es el objetivo que se persigue. Responder a preguntas como *¿qué queremos saber?*, *¿qué datos son necesarios para saberlo?*, *¿cómo se relacionan esos datos entre sí?* son algunas de las preguntas básicas que nos debemos hacer. La razón de ser de cada modelo analítico es el objetivo del proyecto para el que se trabaja.

Existen modelos de varios tipos: **predictivos**, basado en la premisa de si ocurre X pasará Y, **prescriptivos**, ayudan a facilitar la toma de decisiones, y **descriptivos**, utilizados entre otras cosas para establecer relaciones entre los datos. Al final lo que se intenta con todos ellos es construir sistemas que

encuentren patrones útiles en los datos que consigan responder a las hipótesis iniciales planteadas de cada proyecto.

A continuación se resume el procedimiento de manera genérica.

1. El proceso de análisis comienza con la recopilación de datos, en la cual se identifican la información que necesitan para una aplicación de análisis en particular. Es posible que sea necesario combinar los datos de diferentes sistemas de origen a través de rutinas de integración de datos, transformarlos en un formato común y cargarlos en un sistema de análisis. En otros casos, el proceso de recopilación puede consistir en extraer un subconjunto relevante de una secuencia de datos brutos y moverlo a una partición separada en el sistema para que pueda analizarse sin afectar el conjunto de datos completo.
2. Una vez que los datos que se necesitan están guardados, el siguiente paso es encontrar y corregir los problemas de calidad de los datos que podrían afectar la precisión de las aplicaciones de análisis. Eso incluye la ejecución de perfiles de datos y tareas de limpieza de datos para garantizar que la información en un conjunto sea coherente y que se eliminen los errores y las entradas duplicadas.
3. Luego, se realizan trabajos adicionales de preparación para manipular y organizar los datos para el uso analítico planificado, y se aplican políticas de gobierno de datos para garantizar que los datos se ajusten a los estándares corporativos y se usen correctamente.

En otras palabras se podría resumir con lo que se conoce como *proceso ETL*: **Extracción** (obtención de datos de las fuentes de origen), **Transformación** (realización de los cálculos necesarios para obtener los datos que nos interesan) y **Carga** (en esta parte del proceso se vuelcan los datos procedentes de la fase de transformación al sistema de destino.)

Visualización de datos

Una vez tenemos los datos es hora de transformarlos en información. Desde un punto de vista más comercial podríamos afirmar que la visualización de datos es un proceso que consume datos como entrada y los transforma en conocimiento del negocio.

Desde un punto de vista más técnico la visualización de datos es la presentación gráfica de información con dos propósitos principales. Por un lado, la interpretación y construcción de significado a partir de los datos (es decir, el análisis); y por otro lado, la comunicación (la comunicación de la información que se obtiene de esos datos). De una manera más simple se trata de

representar magnitudes de forma visual con el objetivo último de presentar una información. Normalmente esta información se presenta a una audiencia o puede ser también parte de un estudio con el que obtener patrones dentro de un proyecto más grande. En la figura 3.11 se pueden ver los pilares básicos que conforman una buena visualización.

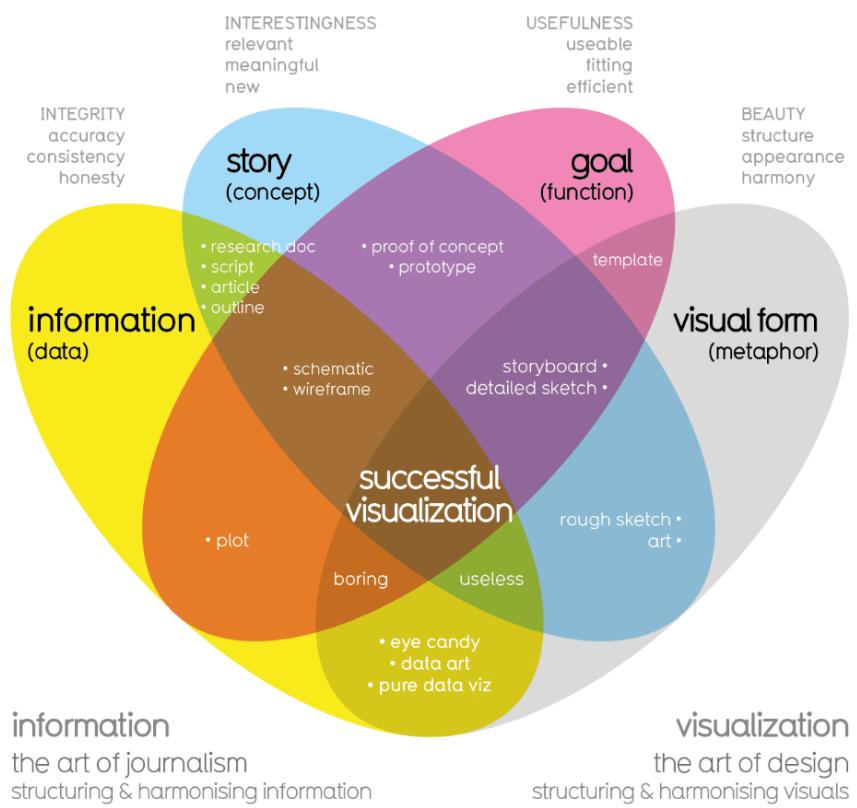


Figura 3.11: *What makes a good visualization?* by David McCandless. Fuente: InformationIsBeautiful.net

Una visualización es una herramienta muy potente para descubrir y comprender la lógica que se encuentra detrás de un conjunto de datos, así como para compartir esta interpretación con otras personas desde un punto de vista objetivo.

El potencial de la visualización de datos es muy muy alto. En muchas ocasiones solemos tener una maraña de datos que por sí solos no nos dicen nada, sin embargo, una vez procesados, transformados y limpios, unos datos pueden convertirse a través de una buena visualización en una información muy relevante.

Algunas organizaciones [17] tiene sus propias guías de estilo para la creación de visualizaciones. Es el caso de por ejemplo Google que tiene un apartado de [Material Design](#) dedicado en exclusiva al *data visualization*.

Técnicas y herramientas

Esta parte de la memoria tiene como objetivo presentar las tecnologías y las herramientas de desarrollo que se han utilizado para llevar a cabo el proyecto. Se han estudiado diferentes alternativas de metodologías, herramientas, y se pretende aquí realizar un resumen de los aspectos más destacados, incluyendo comparativas entre las distintas opciones y, en caso de ser necesario, una pequeña justificación de las elecciones realizadas.

4.1. Metodologías

En este apartado se describen las metodologías utilizadas para el desarrollo del proyecto.

Estrategia de investigación

En el presente proyecto se sigue una estrategia de investigación de análisis de datos cuantitativos. Y dentro de ésta se busca realizar una análisis exploratorio y descriptivo que de pie a sacar una conclusiones de los datos escogidos como punto de partida

“La idea del análisis de datos es buscar patrones en los datos y sacar conclusiones. Existe una amplia gama de técnicas establecidas para analizar datos cuantitativos” [27].

Siguiendo esta estrategia se utilizan técnicas estadísticas exploratorias y descriptivas simples para encontrar patrones en los datos y comprobar si dichos patrones se encuentran realmente en los datos o si son sólo fruto del azar. Además, se utilizan tablas, mapas y gráficos para presentar los datos de una manera visual y sencilla. La recolección de datos se realiza mediante portales de datos abiertos vía descarga ordinaria.

En cuanto a la realización del proyecto respecto de la parte técnica, se hace uso de *Python* como lenguaje de programación de uso general y de *R* como entorno y lenguaje de programación con enfoque estadístico. Además también he usado *SQL* para realizar diversas pruebas. Para los gráficos me apoyo en diferentes herramientas *open source* y para la visualización final en *PowerBi*. También se realizará una pequeña web a modo de resumen para recopilar información relevante del proyecto.

Metodología de trabajo diaria: Scrum

Scrum es un marco de trabajo de desarrollo de software que está dentro de las metodologías ágiles y es una de las más conocidas actualmente.

Se trata de una herramienta muy útil en espacios donde los grupos de trabajo tienen dificultades para hacer las acciones u operaciones que les lleven a objetivos en común. Dicho de otro modo, *Scrum* sirve para que equipos multidisciplinares trabajen en entornos complejos, donde los requisitos son muy cambiantes, y los resultados se tienen que obtener en un plazo corto de tiempo. Se pueden observar las características principales en la figura 4.12.

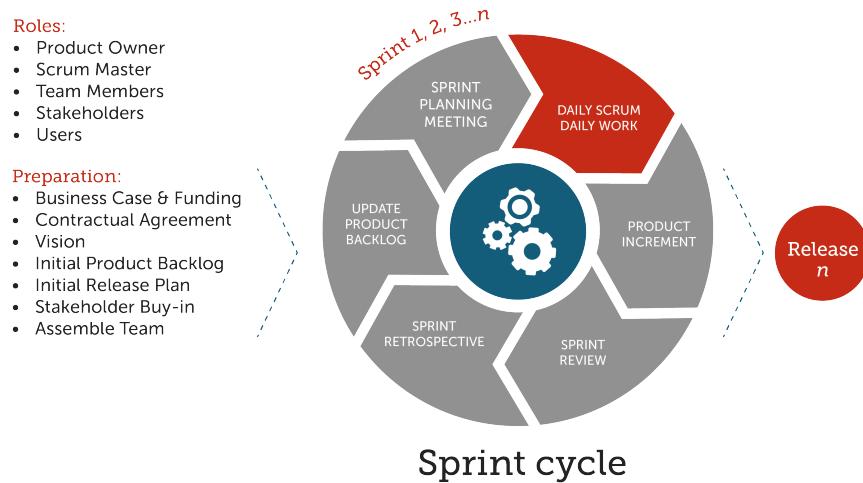


Figura 4.12: Resumen metodología *Scrum*. Fuente: manifesto.co.uk

En resumen, Scrum propone seguir un proceso de desarrollo iterativo e incremental a través de una serie de iteraciones denominadas *sprints* y de revisiones. Hablando en términos personales, obtuve el *título certificado de Scrum Master* el pasado mes de abril y es una metodología que aplico a diario durante el trabajo y que particularmente me apasiona. Si bien es cierto, siguiendo la teoría, que es necesario tener en cuenta que esta metodología fue pensada para trabajar en equipo por lo que en este trabajo se ha intentado

emular de la mejor forma posible dadas las circunstancias: seguimiento de un *board* (ver figura 4.13), reuniones puntuales con los tutores, *sprints* quincenales...

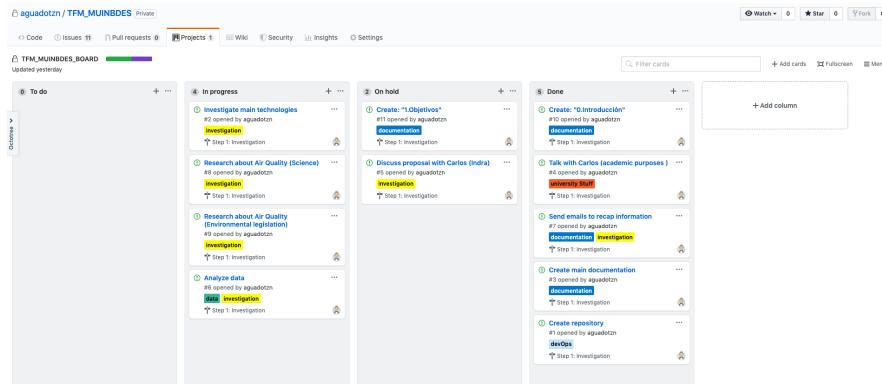


Figura 4.13: *Board* de tareas empleado durante el proyecto. Fuente: Elaboración propia.

Se detallará más la metodología que se ha seguido en este proyecto en el **Apéndice A**.

4.2. Herramientas

En este apartado se describen las herramientas utilizadas para el desarrollo del sistema, concretamente en esta parte hago referencia a la parte software.

Pycharm

[Pycharm](#) es el editor es un IDE (Entorno de desarrollo integrado) desarrollado por la compañía Jetbrains, está basado en *IntelliJ IDEA*. Pycharm tiene cientos de funciones que lo puede ver como una herramienta muy pesada, pero que valen la pena ya que ayuda con el desarrollo del día a día. Es el entorno profesional por excelencia para trabajar con *Python*. Es posible configurar también *jupyter notebooks* para trabajar de maner interna con este editor.

■ Alternativas estudiadas

- [Sublime Text](#): se pensó en esta alternativa debido a que se trata de un IDE bastante liviano en comparación con Pycharm. Finalmente se descartó por aspectos personales.

Jupyter Notebooks

[Jupyter Notebook](#) es un entorno de trabajo interactivo que permite desarrollar código en *Python* de manera dinámica, a la vez que integrar en un mismo documento tanto bloques de código como texto, gráficas o imágenes. Es un *SaaS* utilizado ampliamente en análisis numérico, estadística y *machine learning*, entre otros campos de la informática y las matemáticas.

Conda

[Conda](#) (Anaconda Navigator) es una interfaz gráfica de usuario para el gestor de paquetes y entornos de conda. A través del cuál se ha instalado *Jupyter Notebooks*.

R studio

[R studio](#) es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo. Aunque existen otros, sin duda es el IDE preferido para trabajar con R.

Visual Studio code

[Visual Studio code](#) es un editor de código fuente desarrollado por Microsoft para Windows, Linux y macOS. Incluye soporte para depuración, control de Git embebido, resaltado de sintaxis, finalización inteligente de código, fragmentos y refactorización de código. También es personalizable, para que los usuarios puedan cambiar el tema del editor, los atajos de teclado y las preferencias. Es gratuito y de código abierto, aunque la descarga oficial se realiza bajo licencia propia. Lo mejor de este editor es la gran cantidad de *plugins* que puedes instalar facilitando programar en casi cualquier lenguaje.

Firefox for developers

Para el navegador se ha elegido Firefox en su versión para desarrolladores: [Firefox Developer Edition](#).

4.3. Herramientas de visualización de datos

En este apartado se describen las herramientas utilizadas para el desarrollo la parte de visualización de datos.

Power BI

Power BI es una solución de análisis empresarial que permite visualizar los datos y compartir información con toda la organización, o insertarla en su aplicación o sitio web. En este caso se ha utilizado a través de una máquina virtual windows proporcionada por la universidad.

▪ Alternativas estudiadas

- [Qlik](#), fue visto y utilizado durante el máster y es sin duda una gran alternativa aunque me encuentro más cómodo trabajando con PowerBI.
- [Tableau](#), esta fue nuestra segunda opción. Tableau es una herramienta de visualización muy potente que permite realizar visualizaciones dinámica muy completas. Finalmente fue descartado por problemas de hardware.

CartoDB

CartoDB es una de las empresas líderes en cartografía mundiales. Es un servicio de pago aunque actualmente se ha accedido a él para este trabajo mediante una licencia universitaria que posee github en la que se tiene acceso a diversas herramientas entre las que forma parte **CartoDB**. Ha sido utilizado para la distribución geográfica de las estaciones.

- [Mapbox](#) es una herramienta que permite geolocalizar posiciones entre otras opciones. Es de pago aunque tiene una licencia de uso libre restringida por número de peticiones al servidor. También se ha usado para la visualización final pero se han encontrado contratiempos durante el desarrollo.

Vega

Vega es una herramienta totalmente *open source* que permite, a través de la ingestión de diferentes formatos, crear gráficos interactivos.

- [Voyager2](#) es una herramienta interna de vega que, dado un dataset, permite realizar un análisis exploratorio de las variables que lo forman a través de diversos gráficos.

DataWrapper

DataWrapper es una herramienta que dado un dataset (generalmente ya preprocesado ya que no se trata de una herramienta de limpieza sino que se enfocan más en la parte de visualizado) permite generar gráficos, mapas o tablas

con el objetivo principal de enriquecer contenido en historias escritas. Principalmente está pensado para el acompañamiento de parte gráfica a noticias en diferentes medios de comunicación.

4.4. Tecnologías

En este apartado se describen las tecnologías utilizadas, concretamente lenguajes de programación, para el desarrollo del proyecto.

Python

Python es un lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible. Los programas escritos en Python no necesitan compilarse de antemano para poder ejecutarse, por lo que es fácil probar pequeños fragmentos de código y hacer que el código sea más fácil de mover entre las plataformas.

Este lenguaje tiene licencia de software libre y dispone de una gran comunidad de usuarios que se encarga de enriquecerlo mediante la creación de nuevas librerías, funciones, etc.

Python es muy popular debido a su sencilla integración con otras plataformas y a que dispone de gran cantidad de librerías. Algunas librerías usadas en este proyecto han sido [ggplot](#), [numpy](#), [urllib](#), [pandas](#) o [bokeh](#) entre otras.

R

R es un entorno y un lenguaje de programación enfocado en el análisis estadístico de los más utilizados en el campo de la minería de datos que pueden aplicarse a gran variedad de disciplinas. De nuevo es software libre y es uno de los lenguajes de programación más utilizados en investigación científica.

Algunas librerías usadas en este proyecto han sido [ggplot](#), [numpy](#), [urllib](#), [pandas](#) o [bokeh](#) entre otras.

Librerías R

Tinybird

[Tinybird.co](#) es un **aplicación en beta** que gracias a un cúmulo de circunstancias hemos tenido la suerte de probar en este proyecto. Se trata de una aplicación que utiliza lenguaje *SQL* para realizar consultas sobre datasets. Es decir, dado un *dataset* se crean diversos *pipelines*, esa es la denominación que le dan, con los que podemos analizar, transformar y limpiar los datos de acuerdo a nuestros objetivos. Parece una especie de *notebook* de jupyter aunque no lo es dado que no posee tal cantidad de opciones.

Una vez tenemos los datos preprocesados podemos establecer un *endpoint* a través del cuál es posible acceder de manera externa mediante *API REST*. Después es muy fácil consumirlo desde, por ejemplo, un aplicación web llamando a ese endpoint y accediendo a los datos ya limpios y listos para visualizarlos.

4.5. Documentación

En este apartado se describen algunas de las herramientas utilizadas para la parte de la documentación del proyecto.

LaTeX

LaTeX es un sistema de composición de textos, orientado a la creación de documentos escritos que presenten una alta calidad tipográfica. Por sus características y posibilidades, es usado de forma especialmente intensa en la generación de artículos y libros científicos.

▪ Detalles

- *TeXstudio*: como editor de escritorio de LaTeX se ha utilizado *TeXstudio*, una recomendación personal de Carlos, tutor académico, y que sin lugar a dudas ha supuesto un gran salto de calidad a la hora de realizar el trabajo en cuanto a la parte de la memoria.
- *Tables_generator*: una de las cosas más engorrosas a la hora de usar LaTeX son las tablas. Por esa razón se ha empleado *tables_generator* para realizar esta parte de una manera más cómoda.

Zotero

[Zotero](#) es un gestor de referencias bibliográficas, libre, abierto y gratuito desarrollado por el *Center for History and New Media* de la Universidad George Mason que funciona también como servicio. El funcionamiento de zotero es sencillo y está basado en los principios: recopilar, organizar, citar, sincronizar y colaborar. En la figura 4.14 podemos ver la aplicación de escritorio para Mac.

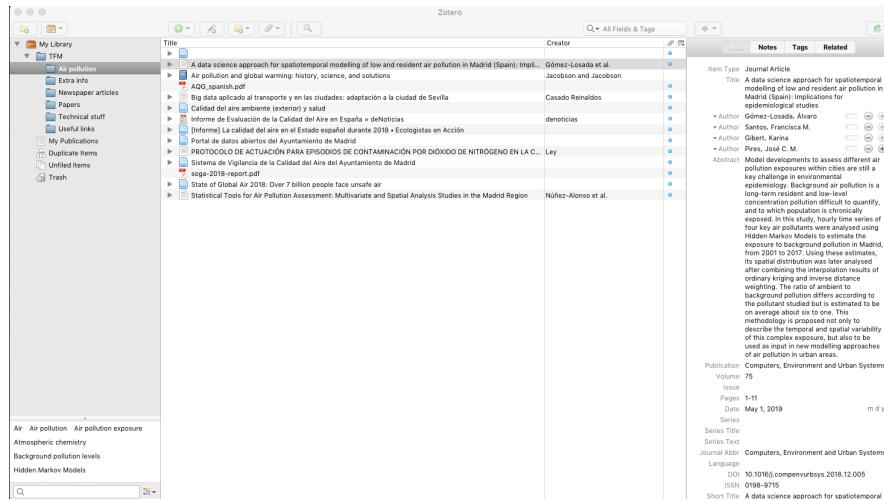


Figura 4.14: Pantalla principal Zotero. Fuente: elaboración propia

▪ Alternativas estudiadas

- **Mendeley** es una aplicación web y de escritorio, propietaria y gratuita. Permite gestionar y compartir referencias bibliográficas y documentos de investigación, encontrar nuevas referencias y documentos y colaborar en línea. Fue descartada debido a problemas en otros trabajos anteriores con la sincronización entre dispositivos.
- **Zenhub** es otro gestor de referencias bibliográficas que descarté por experimentar problemas con la extensión del navegador.

4.6. Otras herramientas

En este apartado se describen otras herramientas que se han empleado también durante el proyecto.

Git

Git [6] es un sistema de control de versiones. Este tipo de sistemas registran los cambios realizados sobre un archivo o conjunto de archivos a lo largo del tiempo, de modo que es posible recuperar versiones específicas del mismo archivo más adelante.

Github

Github es una plataforma cloud de desarrollo colaborativo de software para alojar proyectos utilizando el sistema de control de versiones *git* [6]. El código se almacena de forma pública, aunque también se puede hacer de forma

privada, además desde principios de este mismo año las opciones de repositorios privados se permiten de manera gratuita. GitHub aloja tu repositorio de código y te brinda herramientas muy útiles para el trabajo en equipo.

Se ha utilizado esta herramienta a modo de control de versiones y también se ha usado el *project board* interno como gestor de tareas y de *sprints*. Los *project boards* funcionan, generalmente por repositorio y se pueden utilizar para crear flujos de trabajo personalizados que se adapten a las necesidades de cada proyecto.

▪ Alternativas estudiadas

- [Zenhub](#) es una extensión de Chrome para github. Se utiliza para gestionar proyectos y funciona de manera nativa en la interfaz. Se basa en la metodología ágil y resulta verdaderamente útil a la hora de realizar y gestionar un proyecto.
- [Trello](#) es un gestor de tareas que permite el trabajo de forma colaborativa mediante tableros compuestos de columnas que representan distintos estados. Se basa en el método *Kanban* para gestión de proyectos, con tarjetas que viajan por diferentes listas en función de su estado.

Github pages

Dentro de esas herramientas que se mencionaban al describir Github en el apartado anterior se encuentran [github pages](#). Esta herramienta permite alojar sitios web estáticos gratuitamente dentro un repositorio de código. GitHub pages permite dos modalidades de publicación:

1. La primera es “User site” (solo se podrá tener un sitio de este tipo por cuenta); en este caso el sitio web será publicado en `username.github.io` (siendo `username` el nombre de usuario de la cuenta).
2. La segunda opción es “Project site” (proyectos ilimitados) el cual será publicado en `username.github.io/repository` (siendo `repository` el nombre del repositorio).

Es una manera sencilla, rápida, cómoda de realizar pruebas para una web. Para proyectos de carácter académico en la que mostrar unos resultados en una opción bastante recomendada.

■ Alternativas estudiadas

- Plataformas en la nube: como por ejemplo como Amazon AWS u OpenShift son muy populares hoy en día y permiten prácticamente lo mismo que github pages.
- **Heroku** es un servicio de almacenamiento en la nube que además tiene mecanismos y herramientas para que la puesta en producción de las aplicaciones web sea prácticamente automática.

StackOverFlow

[Stack Over Flow](#) es una de las comunidades de desarrolladores más importantes del mundo en la que se responden cuestiones de diferentes lenguajes. Es una herramienta fundamental para programadores. Actualmente Python es el lenguaje de programación con más consultas en la plataforma.

OpenRefine

[OpenRefine](#) es una potente herramienta *open source* para trabajar con datos: limpiarlos, transformarlos de un formato a otro y ampliarlos con servicios web y/o datos externos. Se utilizó para categorizar y ver los datos en la parte de investigación. En la figura 4.15 una imagen de la herramienta.

	PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANO	MES	D91	V91	D92	V92	D93	V
1.	Facet			Text facet	28079004_1_38	2018	01	00001	V	00001	V	00002	V
2.	Text filter			Numeric facet	28079004_1_38	2018	02	00000	V	00003	V	00002	V
3.				Timeline facet	28079004_1_38	2018	03	00001	V	00002	V	00001	V
4.	Edit cells			Scatterplot facet	28079004_1_38	2018	04	00002	V	00003	V	00002	V
5.	Edit column			Custom text facet...	28079004_1_38	2018	05	00002	V	00002	V	00003	V
6.	Transpose			Custom numeric facet...	28079004_1_38	2018	06	00003	V	00002	V	00002	V
7.	Sort...			Customized facets	28079004_1_38	2018	07	00002	V	00002	V	00002	V
8.	View				28079004_1_38	2018	08	00009	V	00008	V	00008	V
9.					28079004_1_38	2018	09	00010	V	00009	V	00009	V
10.	Reconcile				28079004_1_38	2018	10	00010	V	00011	V	00011	V
11. 28	079	4	1		28079004_1_38	2018	11	00001	V	00048	N	00014	N
12. 28	079	4	1		28079004_1_38	2018	12	00001	V	00012	V	00012	V
13. 28	079	4	6		28079004_6_48	2018	01	0003	V	0003	V	0004	V
14. 28	079	4	6		28079004_6_48	2018	02	0007	V	0004	V	0004	V
15. 28	079	4	6		28079004_6_48	2018	03	0003	V	0005	V	0003	V

Figura 4.15: Detalle vista principal. Fuente: elaboración propia

Aspectos relevantes del desarrollo del proyecto

Este apartado pretende recoger de manera teórica y breve los aspectos más interesantes del desarrollo del proyecto así como su justificación: decisiones que se han tomado, avances evolutivos por tipo y problemas que surgieron durante toda la realización del proyecto.

5.1. Elección del proyecto

Cuando llegó la hora de buscar proyecto, entre las diferentes opciones existían proyectos de carácter más técnico, también orientados a temas más puramente teóricos o de carácter más de investigación. Sin duda, ésta última fue la opción escogida por el reto que esto supone. No existe por lo tanto un ruta exacta a seguir en este tipo de trabajos, sino que se trata más de tener un objetivo concreto pero a sabiendas de que el camino nos va a conducir a sacar unos patrones de los datos que pueden ser sorprendente o incluso inesperados.

5.2. Formación

Toda la formación con la que se ha contado para realizar este trabajo se ha adquirido gracias al máster cuyo proyecto final representa este trabajo. En éste se han aprendido las bases suficientes como para realizar un trabajo de éstas características. Si bien es cierto que la motivación principal era conducir la carrera del alumno hacia el ámbito del *data science* y no ha habido demasiada formación de *R* durante el máster. Es por ello que para esta funcionalidad concreta se han realizado un par de cursos extra. Para concretar algo más los cursos que se han tomado han sido de la plataforma [Open Webinars](#).

5.3. Metodologías/estrategias aplicadas

Ya se han nombrado las metodologías más usadas durante el proyecto: por un lado *Scrum* como metodología de desarrollo iterativa y por otro la estrategia de investigación ha sido la de análisis de datos cuantitativos.

La estrategia de investigación a permitido aliarse perfectamente a la metodología escogida, al desarrollo del código fuente y a la exploración de la parte científica y legislativa del proyecto.

En resumen, ambas decisiones han fomentado de manera positiva tanto la motivación como el progreso del proyecto.

5.4. Desarrollo análisis datos

El modelo empleado para el desarrollo del proyecto comprende la exploración, transformación y posterior visualización de datos, se detalla de manera precisa en el apartado 7.

Hay que destacar aquí que cuando hablamos de Big Data estamos hablando de cantidades realmente enormes pero en este caso, aún con todo los datos descargados de todos los años, no supone una cantidad excesivamente grande de datos. Es por eso que se han elegido este tipo de lenguajes de programación, quizás más enfocados también a otras áreas pero perfectamente válidos para un estudio de este tipo.

[SEGUIR AQUI](#)

5.5. Desarrollo visualización

La parte de visualización comprende varias partes:

- La parte final del *dashboard*, realizada con *PowerBI*.
- Todas las gráficas internas que se han realizado en los diferentes lenguajes de programación, con ayuda de diferentes bibliotecas, y con la finalidad última de explorar los datos o ayudarse a las explicaciones.
- El uso de otras herramientas, ya nombradas, mediante las cuales también se han realizado gráficas. Ya sea de apoyo a la memoria, como explicación añadida web o para demostrar ciertas afirmaciones.

En general ha sido una parte costosa en términos de tiempo. Además se han intentado seguir guías de estilo [12] adecuadas o tener en cuenta gamas de colores aptas realizando la visualización de la mejor manera posible.

5.6. Desarrollo web

Como complemento a la visualización se ha realizado también una pequeña web, alojada en *github pages*, que pretende ser un pequeño resumen vertical de todo lo que se ha estudiado en el proyecto.

5.7. Documentación

La documentación se ha realizado de manera progresiva siempre que se podía añadir partes de la misma y siempre preguntando sobre las cosas que no estaban claras a los tutores, que han ayudado convenientemente a la realización de la misma. Se han realizado dos envíos a los tutores antes del envío final para su corrección .

5.8. Dificultades encontradas

Durante el desarrollo

Una de las partes más importantes del proyecto, en cuestión de tiempo, ha sido la de familiarizarse con los conceptos científicos que requería el mismo.

[SEGUIR AQUI](#)

5.9. Agradecimientos

Me gustaría añadir este pequeño apartado ya que afortunadamente han sido muchas las personas que se han visto involucradas en el proyecto de manera indirecta.

- *Jorge Gomez, Javi Santana* (fundadores de **tinybird**): además del acceso a la *beta* de la aplicación se han prestado a responder de manera altruista todas las preguntas que he realizado sobre la misma.
- *José Amador Viejo* (Director general de control ambiental, **Ayto de Madrid**): ha respondido sin ningún problema a las dudas sobre legislación que se han realizado.
- *Juan Barcena del Rieg* (Portavoz de **Ecologistas en acción**): se han intercambiado varios correos, sobre todo para la parte legislativa y también de la parte ambiental.
- *Javier Di Deco Sampedro* (Data Scientist en **Piperlab**): existe un [bot de twitter](#) realizado por la mencionada empresa que publica en tiempo real los niveles de contaminación por dióxido de nitrógeno de la ciudad

de Madrid. Se envió un mensaje a la empresa para aprender las metodologías empleadas en el desarrollo del *bot* y Javier me atendió de manera amable y eficaz.

- *Brenda Valverde*, actual periodista de **Newtral** mantuvimos una interesante conversación de carácter periodístico sobre contaminación ambiental en general y Madrid central en particular de la que se obtuvieron diversas fuentes para recopilar información. Es posible seguir su cuenta de twitter [aquí](#).
- *Carlos E. Vivaracho Pascual* (tutor académico, **Universidad de Valladolid**): siempre atento durante el desarrollo del proyecto y poniendo las cosas fáciles cuando había incertidumbre.
- *Fernando Cuenca Cabezas* (tutor de empresa, **Minsait**): cómplice de algunas de las decisiones del proyecto siempre se ha prestado a la colaboración durante todo del desarrollo. Además ha resultado fundamental para abordar diversas cuestiones técnicas.

Trabajos relacionados

En esta sección he recopilado algunos recursos que he encontrado útiles y que tratan sobre el mismo tema o temas de temáticas parecidas. Realmente hay muchos estudios sobre la calidad del aire, sobre todo en países como en estados unidos. He encontrado también bastante información y proyectos técnicos, tanto universitarios como proyectos por puro *hobby*, sobre calidad de aire a nivel de la ciudad de Madrid. Quizás el hecho de tener los datos en modo *open data* da a la gente más libertad para usar estos datos y realizar proyectos con ellos.

6.1. Trabajos útiles

- [*AireMadrid 'La realidad de los datos abiertos'*](#) (Ulises Gascón, 2017): Ulises nos habla de las dificultades que supone crear un proyecto *Open Source* que utiliza los datos abiertos de calidad del aire que publica el Ayuntamiento de Madrid para informar y permitir la reutilización efectiva de esa información por parte de los usuarios y otros desarrolladores.
- [*Respira Madrid*](#) (José M. Martínez, 2019): Implementación en Java para el tratamiento e interpretación de datos sobre la calidad del aire en Madrid.
- [*Estudio calidad aire Madrid*](#) (Beatriz@Chucheria, 2016): Análisis sobre la calidad del aire y predicción sobre los niveles de contaminación de la ciudad de Madrid.
- [*Ánálisis de la contaminación y la calidad del aire en Madrid*](#) (Rodrigo de Miguel González, 2016): análisis detallado del NO₂ del aire de la ciudad de Madrid, compuesto que utiliza el Ayuntamiento para poner en marcha los períodos de restricciones por alta contaminación, comprobando el resultado de dichos períodos a lo largo del año.

- *Big data aplicado al transporte y en las ciudades: adaptación a la ciudad de Sevilla* (Alejandro Casado Reinaldos, Trabajo Fin de Grado, Universidad de Sevilla, 2018): estudio analítico de datos de la ciudad de Sevilla.
- *Análisis temporal multivariante de la contaminación atmosférica dentro del distrito metropolitano de Quito* (Juan Luis Manosalvas Paredes, Trabajo Fin de Máster, Universidad Politécnica de Madrid, 2017): se trata de un estudio que tiene similitudes en cuanto a que se realiza un estudio de la contaminación, se presentan las misma partículas y contaminantes, pero resulta curioso ver como varían las legislaciones al tratarse de un país de América Latina.
- *AireMAD* (Fictizia escuela, 2017): aplicación libre desarrollada por Fictizia (escuela de programación) para poder ver los datos de la calidad del Aire de Madrid en tiempo real. Actualmente no se encuentra en desarrollo pero tiene recursos interesante y una robusta documentación disponible [aquí](#).

Entorno experimental

En esta sección se va a exponer de manuscrita el entorno con el que se ha realizado el estudio.

7.1. Esquema

A modo de introducción se han seguido cuatro partes principales en el desarrollo técnico del proyecto en las que se han utilizado diferentes técnicas. Los siguientes esquemas (ver figuras 7.16 y 7.17) describen de manera la línea a seguir en el proyecto.

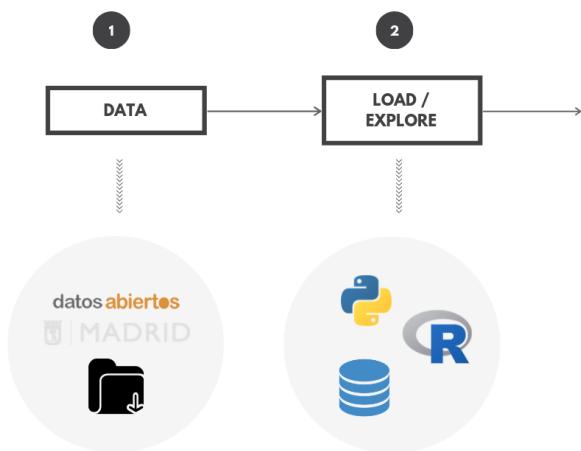


Figura 7.16: Detalle diagrama desarrollo (Partes 1 y 2). Fuente: Elaboración propia.

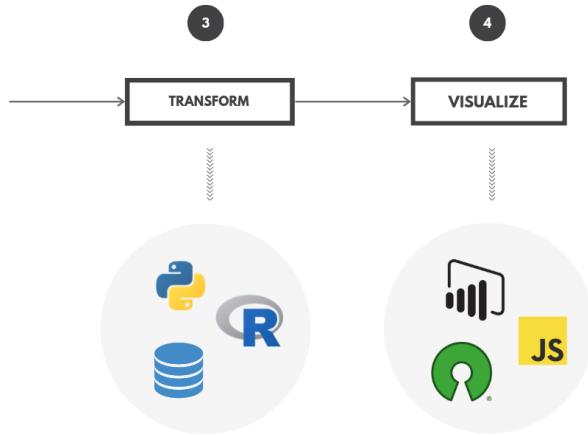


Figura 7.17: Detalle diagrama desarrollo (Partes 3 y 4). Fuente: Elaboración propia.

7.2. Fuente datos origen

En el presente proyecto se han utilizado *datasets* obtenidos principalmente a través de portales de datos abiertos, siendo esta la principal fuente de datos. Estos portales ponen a disposición pública múltiples sets de datos siguiendo una estrategia basada en cuatro aspectos fundamentales: la apertura de datos, la transparencia, la interacción y la participación de personas y/o empresas.

El portal de Datos Abiertos del Ayuntamiento de Madrid, se dedica a promover el acceso a los datos del gobierno municipal y trata de impulsar el desarrollo de herramientas creativas para atraer y servir a la ciudadanía de Madrid. Este portal dispone de un amplio catálogo de datos que puede ser descargado a través de un acceso web para el público general (descarga ordinaria) y un *API REST* con el que automatizar y programar el acceso y descarga de los diferentes *datasets*. A través de este portal [9], es posible descargar un set de datos de calidad del aire con mediciones horarias y diarias desde el año 2001 hasta la actualidad. También existen datos en tiempo real.

Por lo tanto los tres tipos de ficheros que podemos descargar son los siguientes:

- **Datos de calidad del aire horarios:** en este conjunto de datos se puede obtener la información recogida por las estaciones de control de calidad del aire, con los **datos diarios** por anualidades de 2001 a 2019.
- **Datos de calidad del aire horarios:** en este conjunto de datos se puede obtener la información recogida por las estaciones de control de calidad del aire, con los **datos horarios** por anualidades de 2001 a 2019.

Los datos horarios de las magnitudes corresponden a la media aritmética de los valores diezminutales que se registran cada hora.

- **Datos de calidad del aire en tiempo real:** En este conjunto de datos se puede obtener la información actualizada en tiempo real.

En el caso concreto del proyecto nos vamos a centrar en los dos primeros. En el caso de los **datos horarios**, éstos vienen comprimidos en un archivo *.zip* y cuando los descargamos aparecen divididos por meses. El nombre del archivo está formado por las tres letras primeras, que corresponden al mes más un guión bajo seguido de dos números *XX*, que identifican al año.

En caso de los **datos diarios** es similar pero nos proponen el formato de descarga directamente desde la web.

Los principales aspectos a destacar de la fuente de datos serían:

- Los ficheros están en texto plano, *.csv* o *.xml* y los campos no se encuentran delimitados.
- Caso **datos diarios**: los datos se encuentran almacenados mensualmente y agrupados en ficheros comprimidos por cada año.
- Caso **datos horarios**: cada registro contiene los 24 valores horarios de un día, 30 ó 31 filas contiguas corresponden a los valores de los días del mes, repitiéndose con cada magnitud (contaminante) de todas las estaciones que lo miden. Cada fichero contiene un mes de observaciones.
- Se dispone de mediciones de calidad del aire desde el año 2001 hasta la actualidad.
- Cada campo tiene asignado un determinado número de dígitos.
- Todos los campos contienen datos numéricos ya sean identificadores o medidas.
- Hay que precisar que las medidas válidas están marcadas con una V y las medidas no válidas están marcadas con una N. Tan solo serán validas las que contienen la V.
- Los datos anteriores a 2010 presentan algunos errores y los formatos están delimitados

La tabla 7.5 contiene los código de interpretación de los **datos horarios**. Tres apuntes importantes:

1. El campo punto de muestreo incluye el código de la estación completo (provincia, municipio y estación) más la magnitud y la técnica de muestreo.
2. H01 corresponde al dato de la 1 de la mañana de ese día, V01 es el código de validación, H02 al de las 2 de la mañana, V02 y así sucesivamente.
3. Únicamente son válidos los datos que llevan el código de validación “V”.

PROVINCIA	<i>28</i>
MUNICIPIO	<i>79</i>
ESTACION	<i>4</i>
MAGNITUD	<i>1</i>
PUNTO_MUESTREO	<i>28079004_1_38</i>
ANO	<i>2019</i>
MES	<i>1</i>
DIA	<i>1</i>
H01	<i>23</i>
V01	<i>V</i>
H02	<i>17</i>
V02	<i>V</i>
[...]	<i>[...]</i>

Tabla 7.5: Tabla intérprete valores horarios.

La tabla 7.6 contiene los código de interpretación de los **datos diarios**. Tres apuntes importantes:

1. El campo punto de muestreo incluye el código de la estación completo (provincia, municipio y estación) más la magnitud y la técnica de muestreo.
2. D01 corresponde al dato del primer día del mes, D02 al del segundo día y así sucesivamente.
3. Únicamente son válidos los datos que llevan el código de validación “V”.

PROVINCIA	28
MUNICIPIO	79
ESTACION	4
MAGNITUD	1
PUNTO_MUESTREO	28079004_1_38
ANO	2019
MES	1
D01	18
V01	V
D02	20
V02	V
[...]	[...]

Tabla 7.6: Tabla intérprete valores diarios.

Como ya hemos nombrado anteriormente, el Sistema de Vigilancia está formado por 24 estaciones remotas automáticas que recogen la información básica para la vigilancia atmosférica. Poseen los analizadores necesarios para la medida correcta de los niveles de gases y de partículas. La localización de las estaciones de control también se encuentra para su libre descarga, este caso se permiten los ficheros *.xls*, *.csv* y *.geo*. La tabla 7.7 contiene los datos para la correcta interpretación de los ficheros.

Código estación	Lugar	Comentarios
28079001	Pº. Recoletos	Baja.- 04/05/2009 (14:00 h.)
28079002	Glta. de Carlos V	Baja.- 04/12/2006 (11:00 h.)
28079003	Pza. del Carmen	* Código desde enero 2011
28079004	Pza. de España	
28079005	Barrio del Pilar	* Código desde enero 2011
28079006	Pza. Dr. Marañón	Baja.- 27/11/2009 (08:00 h.)
28079007	Pza. M. de Salamanca	Baja.- 30/12/2009 (14:00 h.)
28079008	Escuelas Aguirre	
28079009	Pza. Luca de Tena	Baja.- 07/12/2009 (08:00 h.)
28079010	Cuatro Caminos	* Código desde enero 2011
28079011	Av. Ramón y Cajal	
28079012	Pza. Manuel Becerra	Baja.- 30/12/2009 (14:00 h.)
28079013	Vallecas	* Código desde enero 2011
28079014	Pza. Fdez. Ladreda	Baja.- 02/12/2009 (09:00 h.)
28079015	Pza. Castilla	Baja.- 17/10/2008 (11:00 h.)
28079016	Arturo Soria	
28079017	Villaverde Alto	
28079018	C/ Farolillo	
28079019	Huerta Castañeda	Baja.- 30/12/2009 (13:00 h.)
28079020	Moratalaz	* Código desde enero 2011
28079021	Pza. Cristo Rey	Baja.- 04/12/2009 (14:00 h.)
28079022	Pº. Pontones	Baja.- 20/11/2009 (10:00 h.)
28079023	Final C/ Alcalá	Baja.- 30/12/2009 (14:00 h.)
28079024	Casa de Campo	
28079025	Santa Eugenia	Baja.- 16/11/2009 (10:00 h.)
28079026	Urb. Embajada (Barajas)	Baja.- 11/01/2010 (09:00 h.)
28079027	Barajas	
28079047	Méndez Álvaro	Alta.- 21/12/2009 (00:00 h.)
28079048	Pº. Castellana	Alta.- 01/06/2010 (00:00 h.)
28079049	Retiro	Alta.- 01/01/2010 (00:00 h.)
28079050	Pza. Castilla	Alta.- 08/02/2010 (00:00 h.)
28079054	Ensanche Vallecas	Alta.- 11/12/2009 (00:00 h.)
28079055	Urb. Embajada (Barajas)	Alta.- 20/01/2010 (15:00 h.)
28079056	Plaza Elíptica	Alta.- 18/01/2010 (12:00 h.)
28079057	Sanchinarro	Alta.- 24/11/2009 (00:00 h.)
28079058	El Pardo	Alta.- 30/11/2009 (13:00 h.)
28079059	Parque Juan Carlos I	Alta.- 14/12/2009 (00:00 h.)
28079086	Tres Olivos	Alta.- 14/01/2010 (13:00 h.)

Tabla 7.7: Tabla intérprete estaciones de control.

Y para terminar los contaminantes, descritos en la tabla 7.8, también es necesario interpretarlos.

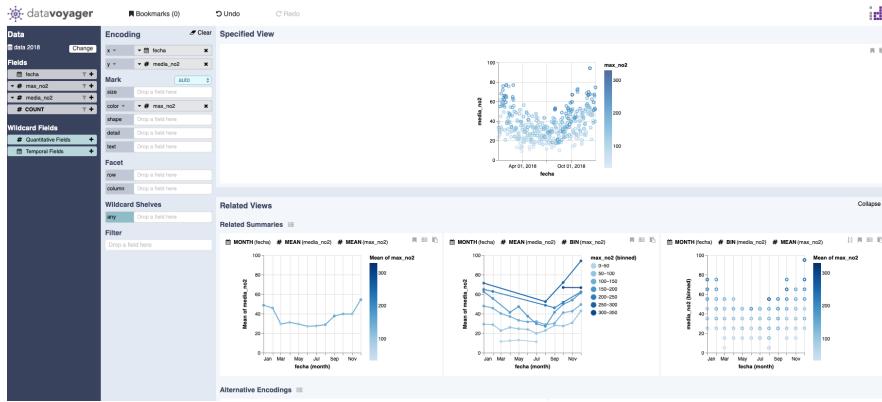
	Magnitud	Abreviatura	Unidad	Técnica de medida	
01	<i>Dióxido de Azufre</i>	SO ₂	µg/m ³	38	Fluorescencia ultravioleta
06	<i>Monóxido de Carbono</i>	CO	mg/m ³	48	Absorción infrarroja
08	<i>Dióxido de Nitrógeno</i>	NO ₂	µg/m ³	48	Id.
07	<i>Monóxido de Nitrógeno</i>	NO	µg/m ³	08	Quimioluminiscencia
09	<i>Partículas <2.5 µm</i>	PM _{2,5}	µg/m ³	47	Microbalanza
10	<i>Partículas <10 µm</i>	PM ₁₀	µg/m ³	47	Id.
12	<i>Óxidos de Nitrógeno</i>	NO _x	µg/m ³	08	Quimioluminiscencia
20	<i>Tolueno</i>	TOL	µg/m ³	59	Gases
30	<i>Benceno</i>	BEN	µg/m ³	59	Id.
35	<i>Etilbenceno</i>	EBE	µg/m ³	59	Id.
37	<i>Metaxileno</i>	MXY	µg/m ³	59	Id.
38	<i>Paraxileno</i>	PXY	µg/m ³	59	Id.
39	<i>Ortoxileno</i>	OXY	µg/m ³	59	Id.
42	<i>Hidrocarburos totales(hexano)</i>	TCH	mg/m ³	02	Ionización de llama
43	<i>Metano</i>	CH ₄	mg/m ³	02	Id.
44	Hidrocarburos no metánicos	NMHC	mg/m ³	02	Id.

Tabla 7.8: Tabla ínterprete contaminantes.

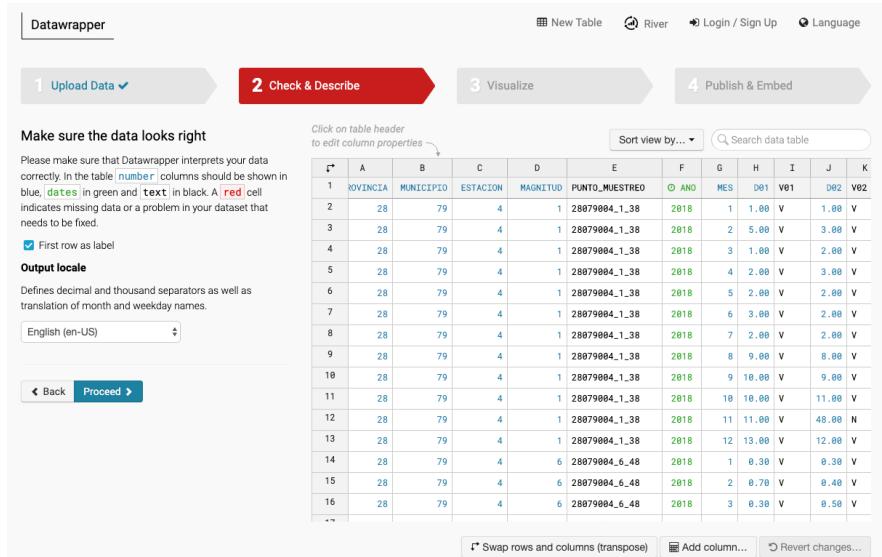
7.3. Exploración

Tras conocer, estudiar a fondo los tipos de datos existentes y su correcta interpretación es tiempo de explorarlos. La fase de exploración se ha realizado de diferentes maneras y desde varias perspectivas con la ayuda de diversas herramientas *open source*.

Esta fase es necesaria para realizar una investigación preliminar con el fin de entender mejor las características específicas de nuestros datos. En esta fase buscaremos correlaciones, tendencias y valores atípicos. Sin esta fase, no podríamos utilizar los datos de manera eficaz. Hemos hecho uso de diversas herramientas como es el caso de [Voyager2](#) (figura 7.18).

Figura 7.18: Detalle *vega-voyager*. Fuente: Elaboración propia.

Datawrapper (ver figura 7.19) es otra herramienta que nos ayuda a analizar nuestros datos de una manera rápida. Es conveniente no coger una muestra demasiado grande ya que, ni es necesario, ni el software está correctamente optimizado para grandes volúmenes.

Figura 7.19: Detalle *Datawrapper*. Fuente: Elaboración propia.

A continuaciónn, en las tablas 7.9 y 7.9, se expone la tipología de las variables en su conjunto. Aunque es posible que no se utilicen todas, sí es necesario introducir aquí todas las variables para establecer una relación con lo descrito en páginas anteriores.

PROVINCIA	<i>integer</i>
MUNICIPIO	<i>integer</i>
ESTACION	<i>integer</i>
MAGNITUD	<i>integer</i>
PUNTO_MUESTREO	<i>integer</i>
ANO	<i>integer</i>
MES	<i>integer</i>
DIA	<i>integer</i>
H01	<i>integer</i>
V01	<i>float</i>
H02	<i>integer</i>
V02	<i>text</i>
[...]	<i>[...]</i>

Tabla 7.9: Tabla tipo datos valores horarios.

PROVINCIA	<i>integer</i>
MUNICIPIO	<i>integer</i>
ESTACION	<i>integer</i>
MAGNITUD	<i>integer</i>
PUNTO_MUESTREO	<i>integer</i>
ANO	<i>integer</i>
MES	<i>integer</i>
D01	<i>integer</i>
V01	<i>text</i>
D02	<i>integer</i>
V02	<i>text</i>
[...]	<i>[...]</i>

Tabla 7.10: Tabla tipo datos valores diarios.

Para conocer más a fondo nuestros datos algunas estadísticas básicas de este tipo que debemos calcular para nuestro conjunto de datos son la media, mediana, el rango y la desviación estándar. La media y la mediana son medidas de la ubicación de un conjunto de valores. La moda es el valor que ocurre con mayor frecuencia en el conjunto de datos. El rango o la desviación estándar son medidas de la dispersión de los datos. Examinar estas mediciones nos dará una idea más precisa de la naturaleza de nuestros datos.

Este parrafo de arriba QUE EH

7.4. Transformación de datos

Tras explorar y entender los datos es hora de transformarlos de acuerdo a nuestros objetivos finales. Es cierto que se han utilizado diferentes lenguajes de programación pero en todos se han seguido patrones comunes común a la hora de transformar los datos.

Algunas características comunes de la transformación de los datos han sido el renombramiento de columnas, el establecer el tipo de dato adecuado por columnas para realizar operaciones y evitar problemas a la hora de realizar operaciones, el intercambiar las columnas de validación para establecer los datos que son correctos o el agrupar por contenido para obtener una determinada gráfica.

Este parrafo de arriba QUE EH

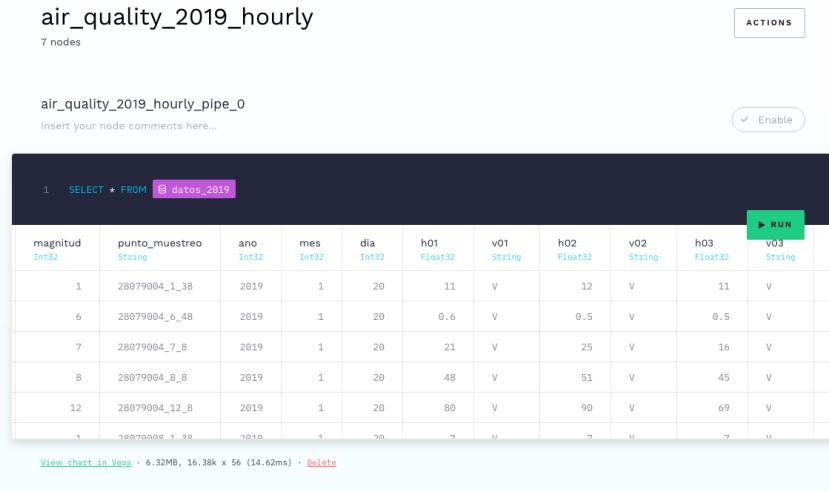
Vamos a poner un ejemplo del procedimiento a seguir para [Tinybird.co](#). Lo primero que hay que hacer es cargar los datos. Como hemos nombrado se trata de una aplicación en beta por lo tanto el entorno gráfico no está disponible todavía para subir cantidades grandes de datos por lo que he creado un script en bash que a través de un bucle sube todos los archivos.

```
#!/bin/bash

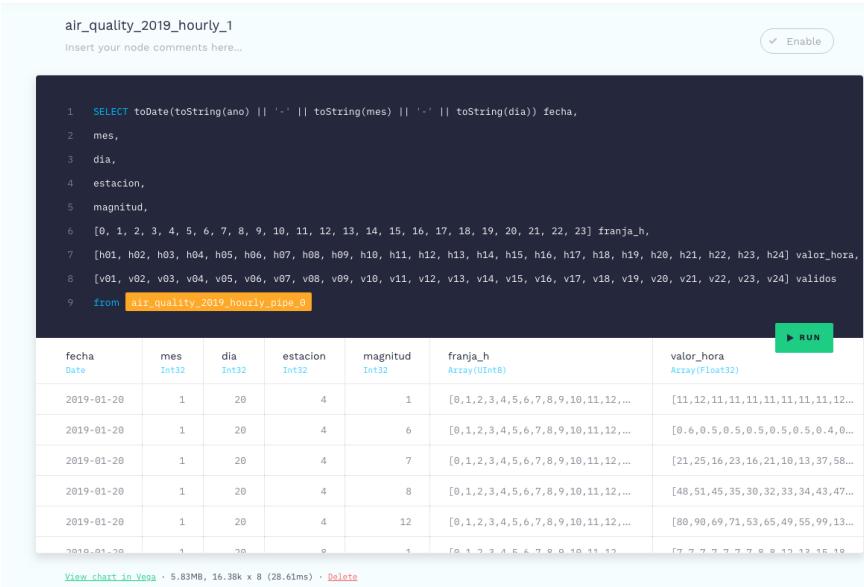
#Personal Token
TOKEN=p eyXXXXXXXXXXXXXXXXXXXXXX
#URL
BASE_URL=/Users/xxxx/Documents//TFM_MUINBDES/Code/data

#Load months
for month in ene feb mar abr may jun jul ago sep oct nov dic
do
curl -F "csv=@${month}_mo18.csv" \
-H "Authorization: Bearer $TOKEN" \
-X POST "https://XXXXXXXXXX/v0/datasources?name=datos_2018&mode=
append"
done
```

Una vez tenemos los datos cargados (ver figura 7.20) es el momento de transformar los datos.

Figura 7.20: Datos cargados en *tinybird*. Fuente: Elaboración propia.

En este caso queremos observar si existen patrones anuales en los datos de la cantidad de CO₂ para después comprobar la misma hipótesis de manera semanal y por horas. Lo primero que hacemos es seleccionar de nuestro conjunto de datos lo que nos hace falta, además transformamos la fecha (ya que recordamos que la fecha viene en nuestro fichero de datos original separada en tres columnas diferentes) para una mejor visualización (ver figura 7.21).

Figura 7.21: Detalle *tinybird*. Fuente: Elaboración propia.

Después escogemos los valores válidos y también la magnitud que corresponde al contaminante del cuál queremos hacer el análisis (ver figura 7.22)

The screenshot shows a data visualization tool interface. At the top, there is a code editor with the following SQL query:

```
1 SELECT * FROM air_quality_2019_hourly_1 array join franja_h,valor_hora,validos where validos = 'V' and magnitud = 8
```

Below the code editor is a table with the following columns: fecha (Date), mes (Int32), dia (Int32), estacion (Int32), magnitud (Int32), franja_h (UInt8), valor_hora (Float32), and validos (String). The table contains 8 rows of data. A green 'RUN' button is located at the bottom right of the table.

fecha Date	mes Int32	dia Int32	estacion Int32	magnitud Int32	franja_h UInt8	valor_hora Float32	validos String
2019-01-20	1	20	4	8	1	48	V
2019-01-20	1	20	4	8	2	51	V
2019-01-20	1	20	4	8	3	45	V
2019-01-20	1	20	4	8	4	35	V
2019-01-20	1	20	4	8	5	30	V
2019-01-20	1	20	4	8	6	29	V
2019-01-20	1	20	4	8	7	29	V
2019-01-20	1	20	4	8	8	29	V

At the bottom left, there is a link to 'View chart in Vega' and a note about file size and execution time.

Figura 7.22: Detalle *tinybird*. Fuente: Elaboración propia.

Para terminar agrupamos y ordenamos por lo que nos hace falta para nuestro análisis (ver figura 7.23) y ya podemos exportar el archivo para proceder a su visualización.

The screenshot shows a data visualization tool interface. At the top, there is a code editor with the following SQL query:

```
1 SELECT fecha dia_semana, valor_hora valor_co2 FROM air_quality_2019_hourly_1
2 WHERE mes == 7
3 GROUP BY valor_co2, dia_semana
4 ORDER BY dia_semana
```

Below the code editor is a table with the following columns: dia_semana (Date) and valor_co2 (Float32). The table contains 6 rows of data. A green 'RUN' button is located at the bottom right of the table.

dia_semana Date	valor_co2 Float32
2019-07-01	81
2019-07-01	38
2019-07-01	74
2019-07-01	24
2019-07-01	63
2019-07-01	25

At the bottom left, there is a link to 'View chart in Vega' and a note about file size and execution time.

Figura 7.23: Detalle *tinybird*. Fuente: Elaboración propia.

Explicar python

7.5. Visualización de datos

Una vez que tenemos los datos transformados de manera adecuada ya podemos concentrarnos en la parte visual. En este trabajo la parte visual está

dividía en varias partes, por un lado en todas las gráficas que hemos empleado para explorar los datos; segundo, en una visualización final a modo de dashboard realizada con *PowerBi*; y en una pequeña web resumen de resultados.

Siguiendo con el ejemplo nombrado en el apartado anterior y cargando los datos en vega podemos obtener la visualización de acuerdo a nuestro objetivo. En este caso hemos calculado la media de NO₂ anual junto con los máximos (ver figura ??) repartido por meses.

Aquí poner datos actualizados de 2019

También lo mismo pero por día de la semana (ver figura 7.24), en este caso par el mes de julio.

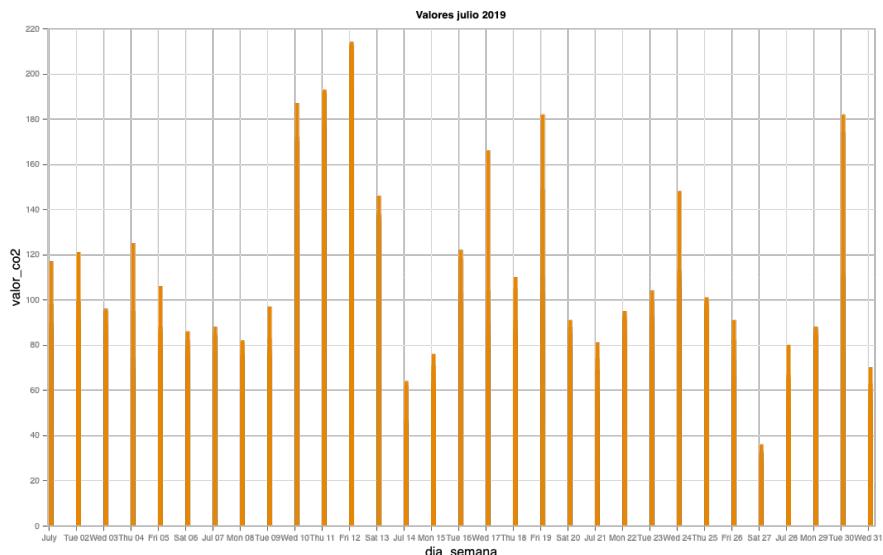


Figura 7.24: Detalle NO₂ 2019 (julio). Fuente: Elaboración propia.

Y lo mismo pero por horas, en este caso para el 12 de julio (ver figura 7.25).

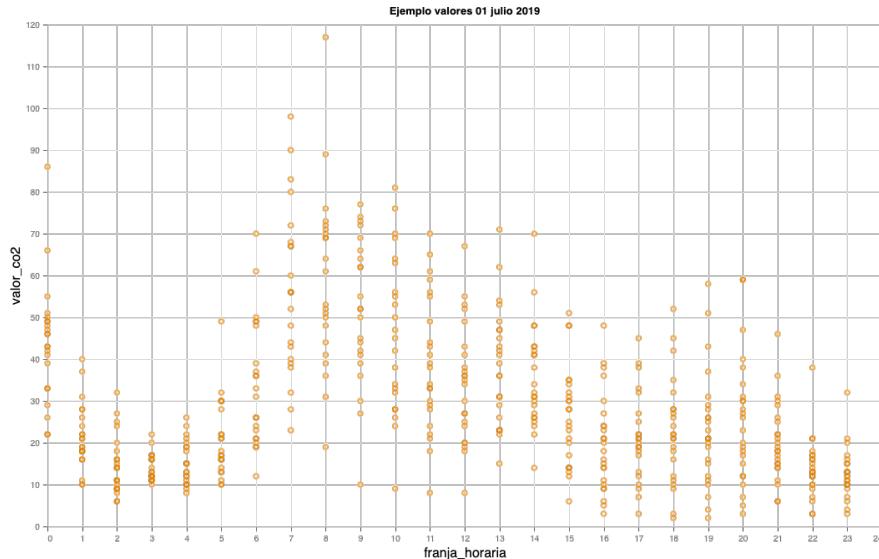


Figura 7.25: Detalle NO₂ 2019 (12 julio). Fuente: Elaboración propia.

La parte de la visualización en **PowerBi** es más interactiva y resulta muy sencilla de utilizar. Una vez transformados los datos mediante python (descrito en el apartado anterior) se cargan los datos. En este caso he cargado los datos de 2014 a 2018, ya que he considerado que terminar el período anual completo era mejor para el análisis.

He realizado distintas páginas dentro de PowerBi en las que podemos ver diferentes tipos de gráficos.

- Página 1:
- Página 2:



Figura 7.26: Detalle page en PowerBi carga. Fuente: Elaboración propia.

■ Página 3:

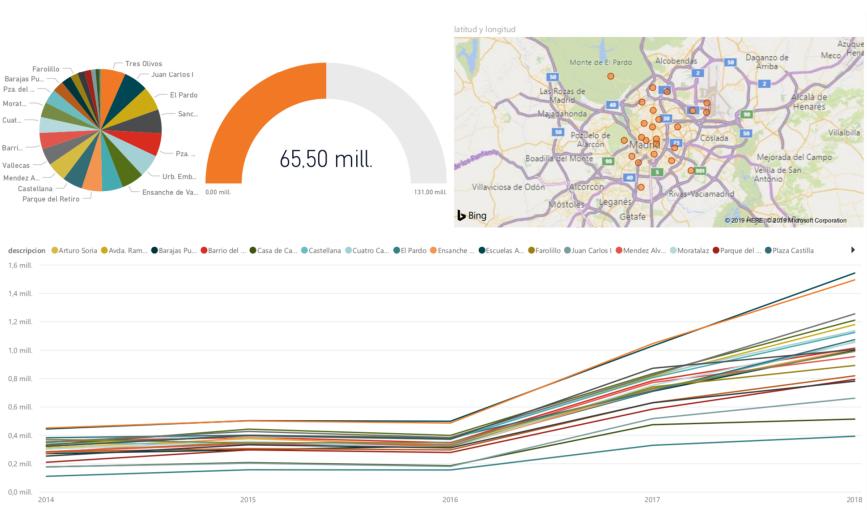


Figura 7.27: Detalle page en PowerBi carga. Fuente: Elaboración propia.

■ Página 4:

Después también se ha experimentado con otro tipo de gráficos en los que se puede ver de manera clara según el año, figura 7.28, cual es la zona más afectada, o en el conjunto completo del período seleccionado (2014-2018) cuál es la zona más afectada, figura 7.29.



Figura 7.28: Ejemplo gráfico por tamaño. Fuente: Elaboración propia.



Figura 7.29: Ejemplo gráfico nube de palabras.

Resultados

En esta penúltima sección se exponen las conclusiones derivadas del trabajo.

8.1. Preámbulo

Los resultados son el remate final a este estudio, sin embargo, quiero hacer aquí una pequeña aclaración. Los efectos de este estudio realizado por otro conjunto de personas y partiendo de los mismos datos pueden ser diferentes dado que el enfoque que se le da es distinto. Por ejemplo yo me he enfocado más en un contaminante, NO₂, y en éste comparado a los niveles de la OMS. Si bien es cierto que se ha encontrado patrones similares con otros trabajos (ver más en la sección 6) no era este el objetivo final.

En la figura 8.30 se puede ver el conjunto de desarrollo completo que se ha seguido para llegar a las visualizaciones.

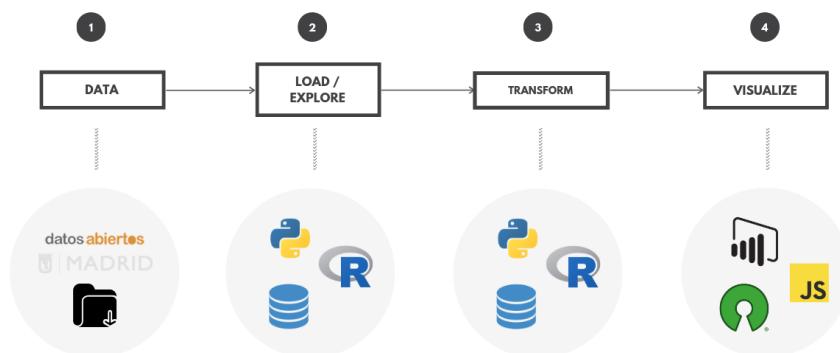


Figura 8.30: Diagrama desarrollo. Fuente: Elaboración propia.

Los resultados que se exponen en el siguiente apartado son los que se han obtenido de observar, analizar y estudiar esas visualizaciones para dar respuesta a las preguntas analíticas.

8.2. Resultados generales obtenidos

Los resultados a los que hemos llegados son los siguientes:

- La contaminación

8.3. Preguntas analíticas

Algunas de las preguntas analíticas, marcadas al inicio del proyecto, han sido las siguientes:

- La contaminación
- Los protocolos de anticontaminación se basan en los niveles de los óxidos de nitrógeno. Estos niveles suben por la noche por lo que siempre se espera a última hora para activar dichos protocolos. Es una afirmación correcta ?
- Existen contaminantes que puedan estar asociados a temporalidad ?
- Consideras que las últimas medidas aplicadas por el Ayuntamiento de Madrid durante los últimos años han mejorado la calidad del aire? Que contaminantes? Alguna zona/estación en concreto ? Puedes confirmar esta información contrastando la evolución temporal del número de alertas, avisos y preavisos en las diferentes zonas ?
- Qué área es la más contaminada ?

Acuérdate de por qué hemos empezado primero reuniendo los datos y analizándolos: para buscar conocimientos útiles o valiosos dentro de todos estos conjuntos de datos, para responder a preguntas o para mejorar los procesos de negocio. Por ejemplo, ¿Debemos cambiar algo en nuestro proceso para eliminar cuellos de botella?, ¿Deberíamos añadir datos a la aplicación para que sea más precisa?, ¿Debemos segmentar nuestra población en grupos mejor definidos para tener un marketing dirigido más eficaz? Este es el primer paso para convertir el conocimiento en acción.

Una vez que hayamos decidido como actuar, el siguiente paso es averiguar como implementar la acción. ¿Que es necesario para añadir esta acción a nuestro proceso o aplicación? ¿Como vamos a automatizarla? Debemos identificar a los grupos de interés y hacer que se involucren en este cambio.

Al igual que sucede con cualquier optimización de procesos, tenemos que monitorizar y medir el impacto de la acción en el proceso o aplicación. Evaluar el impacto conlleva una evaluación de resultados.

La evaluación de resultados de la acción aplicada determinará los pasos a seguir: ¿Necesitamos llevar a cabo un análisis adicional con el fin de obtener mejores resultados? ¿Qué datos debemos revisar? ¿Qué posibilidades adicionales debemos investigar? Por ejemplo, no olvidemos lo que nos permite hacer Big Data: acciones en tiempo real basadas en la transmisión de flujos de información a alta velocidad. Tenemos que definir qué parte del negocio necesita acciones en tiempo real para poder influir en las operaciones o en la interacción con el cliente. Una vez que hayamos definido estas acciones en tiempo real, tenemos que asegurarnos de que existan sistemas automatizados o procesos para la realización de dichas acciones y proporcionar mecanismos de recuperación ante fallos en caso de problemas.

Conclusiones y Líneas de trabajo futuras

En esta última sección se exponen las conclusiones finales derivadas del trabajo. De la misma manera abordaremos las líneas de trabajo futuro que considero que se pueden o deberían realizarse.

9.1. Conclusiones proyecto

Las conclusiones alcanzadas.

- HOla

9.2. Conclusiones personales

Me gustaría exponer también unas conclusiones personales a modo de lista sobre lo que este trabajo ha supuesto para mí.

9.3. Líneas de trabajo futuras

Es importante remarcar las líneas de futuro que el proyecto pretende seguir.

- Refactorizar

Apéndices

Bibliografía

- [1] Calidad del aire ambiente (exterior) y salud. URL [https://www.who.int/es/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health). (Citado en páginas 25 y 26.)
- [2] Sistema de vigilancia de la calidad del aire del ayuntamiento de madrid. URL <http://www.mambiente.madrid.es/sica/scripts/index.php>. (Citado en páginas VII y 14.)
- [3] ¿cómo serán las ciudades del futuro? - BBC news mundo. URL https://www.bbc.com/mundo/noticias/2013/02/130218_ciudades_futuro_ap. (Citado en página 2.)
- [4] Big data para ciudades inteligentes, los datos para mejorar los servicios en las ciudades. URL <https://blog.ferrovial.com/es/2017/10/big-data-ciudades-inteligentes-servicios/>. (Citado en página 9.)
- [5] Formatos - portal de datos abiertos del ayuntamiento de madrid. URL <https://datos.madrid.es/portal/site/egob/menuitem.400a817358ce98c34e937436a8a409a0/?vgnextoid=b07512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextchannel=b07512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>. (Citado en página 29.)
- [6] Git. URL <https://en.wikipedia.org/w/index.php?title=Git&oldid=906157860>. Page Version ID: 906157860. (Citado en página 40.)
- [7] [informe] la calidad del aire en el estado español durante 2018 • ecologistas en acción. URL <https://www.ecologistasenaccion.org/96516/informe-la-calidad-del-aire-en-el-estado-espanol-2018/>. (Citado en páginas V, 11, 15 y 20.)
- [8] Report: London no safer for all its CCTV cameras. ISSN 0882-7729. URL <https://www.csmonitor.com/World/Europe/2012/0222/>

- Report-London-no-safer-for-all-its-CCTV-cameras. (Citado en página 9.)
- [9] Portal de datos abiertos del ayuntamiento de madrid. URL <https://datos.madrid.es/portal/site/egob/>. (Citado en páginas v, 21 y 50.)
- [10] State of global air 2018: Over 7 billion people face unsafe air. URL <https://www.healtheffects.org/announcements/state-global-air-2018-over-7-billion-people-face-unsafe-air>. (Citado en página 25.)
- [11] Universidad carlos III de madrid - la inteligencia ubicua cambiará nuestra vida en menos de una década. URL http://portal.uc3m.es/portal/page/portal/actualidad_cientifica/noticias/inteligencia_ubicua. (Citado en página 10.)
- [12] All the 'little of visualisation of design'. URL <https://www.visualisingdata.com/2016/03/little-visualisation-design/>. (Citado en página 44.)
- [13] Wayback machine. URL https://web.archive.org/web/20120414062644/http://www.magrama.gob.es/imagenes/es/PNMCA_tcm7-181205.pdf. (Citado en página 24.)
- [14] Apache-Github. Apache license 2.0, 2004. URL <https://choosealicense.com/licenses/apache-2.0/>. (Citado en página 83.)
- [15] Ben Balter. Open source license usage on github.com, 2015. URL <https://github.com/blog/1964-open-source-license-usage-on-github-com>. (Citado en página 83.)
- [16] Alejandro Casado Reinaldos. Big data aplicado al transporte y en las ciudades: adaptación a la ciudad de sevilla. URL <https://idus.us.es/xmlui/handle/11441/79499>. (Citado en página 17.)
- [17] Amy Cesal. What are data visualization style guidelines? URL <https://medium.com/nightingale/style-guidelines-92ebe166addc>. (Citado en página 32.)
- [18] denoticias. Informe de evaluación de la calidad del aire en españa » deNoticias. URL <https://www.denoticias.es/notas/informe-de-evaluacion-de-la-calidad-del-aire-en-espana.html>. (Citado en páginas v, VII, VII, VII, 10, 15, 16, 18, 19 y 26.)
- [19] Ecologistas en Acción. [informe] la calidad del aire en el estado español durante 2018. URL <https://www.ecologistasenaccion.org/96516/>

- [informe-la-calidad-del-aire-en-el-estado-espanol-2018/](https://www.dataversity.net/informe-la-calidad-del-aire-en-el-estado-espanol-2018/). (Citado en página 2.)
- [20] Keith D. Foote. Big data vs. smart data. URL <https://www.dataversity.net/big-data-vs-smart-data/>. (Citado en página 8.)
- [21] Github. Choose an open source license, 2016. URL <https://choosealicense.com/>. (Citado en página 83.)
- [22] Álvaro Gómez-Losada, Francisca M. Santos, Karina Gibert, and José C. M. Pires. A data science approach for spatiotemporal modelling of low and resident air pollution in madrid (spain): Implications for epidemiological studies. 75:1–11. ISSN 0198-9715. doi: 10.1016/j.compenvurbssys.2018.12.005. URL <http://www.sciencedirect.com/science/article/pii/S0198971518304447>. (Citado en página 20.)
- [23] Philippe B Laval. Introduction to the mathematics of big data. page 8. (Citado en página 6.)
- [24] La Ley. PROTOCOLO DE ACTUACIÓN PARA EPISODIOS DE CONTAMINACIÓN POR DIÓXIDO DE NITRÓGENO EN LA CIUDAD DE MADRID. page 20. (Citado en página 27.)
- [25] MIT-Github. Mit license, 2017. URL <https://choosealicense.com/licenses/mit/>. (Citado en página 83.)
- [26] David Núñez-Alonso, Luis Vicente Pérez-Arribas, Sadia Manzoor, and Jorge O. Cáceres. Statistical tools for air pollution assessment: Multivariate and spatial analysis studies in the madrid region. 2019:1–9. ISSN 2090-8865, 2090-8873. doi: 10.1155/2019/9753927. URL <https://www.hindawi.com/journals/jamc/2019/9753927/>. (Citado en página 20.)
- [27] B.J. Oates. *Researching information systems and computing*. SAGE. ISBN 978-1-4129-0223-6. URL <https://books.google.es/books?id=1JAeAQAAIAAJ>. (Citado en página 33.)
- [28] Yakov Shafranovich {\textless}ietf@shaftek.org{\textgreater}. Common format and MIME type for comma-separated values (CSV) files. URL <https://tools.ietf.org/html/rfc4180>. (Citado en página 29.)
- [29] Anirudh Sharma. Anirudh sharma: Ink made of air pollution | TED talk. URL https://www.ted.com/talks/anirudh_sharma_ink_made_out_of_air_pollution/transcript-8889. (Citado en página 2.)

Apéndice A

Plan de Proyecto Software

A.1. Introducción

En este capítulo se detalla la planificación del proyecto.

La planificación temporal del proyecto, o lo que es lo mismo, la elaboración del calendario o programa de tiempos, consiste en una representación gráfica de todas las actividades del proyecto necesarias para producir el resultado final que se desea.

Se han utilizado metodologías ágiles para el desarrollo del proyecto y de este modo, se ha realizado un desarrollo dividido en iteraciones. Terminada una iteración empezaba la siguiente y se agregaban a las tareas planeadas las que no habían sido completado de la iteración precedente. Las iteraciones del proyecto, los sprints, estaban pensadas para durar unos **quince días** aproximadamente. No obstante, hay alguna excepción en la que la iteración duró más tiempo, o menos. También existe alguna demora entre algún sprint debido a que tenía demasiada carga de las asignaturas o demasiada carga laboral.

A.2. Planificación temporal

Desde el comienzo del proyecto se planteó utilizar una metodología ágil como *Scrum* para la gestión del proyecto. Aunque no se ha seguido al 100 % la metodología al tratarse de un proyecto académico, sí que se ha aplicado en líneas generales una filosofía ágil y metódica. La diferencia fundamental radica en que esta metodología esta pensada para equipos y no para individuos.

Desarrollo con Scrum

Scrum es un marco de trabajo que define un conjunto de prácticas y roles, y que puede tomarse como punto de partida para definir el proceso de desa-

rrollo que se ejecutará durante el proyecto. Indicar aquí que se ha utilizado terminología en inglés dado que por convenio se siguen estos términos en cualquier organización. De la misma manera las tareas también se han descrito en inglés.

Recordemos que Scrum no define como tal un método o herramienta para llevar el seguimiento del trabajo realizado. Scrum solo propone una serie de buenas prácticas por lo que el seguimiento se puede realizar con una simple hoja de excell, cuaderno o con herramientas de gestión especializadas. Yo he elegido el *project board* de github dentro de mi repositorio del trabajo fin de máster.

Los roles o actores principales en Scrum serían los siguientes:

- *Scrum Master*: su trabajo prioritario es eliminar los obstáculos que impiden que el equipo no alcance el objetivo del *sprint*. Esta persona no es el líder del equipo sino el nexo entre todos ellos. El autor del trabajo asume este rol.
- *Product Owner*: representa a los *stakeholders*¹, se asegura de que el equipo trabaje de forma adecuada desde la perspectiva de éstos. En un entorno real sería el encargado de escribir historias de usuario, priorizarlas y colocarla en el *Product Backlog*. En este caso que se describe, recordemos en un entorno académico, el tutor académico sería el más indicado para asumir este rol.
- *Team* (equipo): en un entorno real sería el equipo que conforman todos los profesionales de diferentes ramas de conocimiento. El objetivo principal es el de entregar el producto. En este caso el autor asume el rol de equipo.

A continuación se describe el ciclo de desarrollo. Al inicio del proyecto el *Product Owner* debe definir los requisitos que serán los objetivos a cumplir. Éstos quedan reflejados y ordenados por orden de importancia en el *Product Backlog*. Durante los sprints cada uno de estos objetivos serán subdivididos en tareas más pequeñas. En la figura A.1 se puede ver un ejemplo del *Backlog*.

¹En toda organización, además de sus propietarios, participan diversos actores claves y grupos sociales que están constituidos por las personas o entes que, de una manera y otra, tienen interés en el desempeño de una empresa porque están relacionadas, bien directa, bien indirectamente, con ella. Fuente Wikipedia. [https://es.wikipedia.org/wiki/Parte_interesada_\(empresas\)](https://es.wikipedia.org/wiki/Parte_interesada_(empresas))

		Author	Labels	Projects	Milestones	Assignee	Sort
14 Open ✓ 17 Closed							
② Create "3.Conceptos teóricos" (Part IV: theoretical introduction to technical concepts) documentation	documentation						
#19 by aguadotzn was closed 8 days ago	Sprint 2: Main ...						
② Create scripts. data data-processing							
#16 by aguadotzn was closed 7 days ago	Sprint 2: Main ...						
② Preprocess data data data-processing							
#15 by aguadotzn was closed 7 days ago	Sprint 2: Main ...						
② Create "3.Conceptos teóricos" (Part III: Legislative analysis) documentation	documentation						
#14 by aguadotzn was closed 9 days ago	Sprint 2: Main ...						
② Create "3.Conceptos teóricos" (Part II: Scientific analysis) documentation	documentation						
#13 by aguadotzn was closed 9 days ago	Sprint 2: Main ...						
② Create "3.Conceptos teóricos" (Part I: Big Data) documentation	documentation						
#12 by aguadotzn was closed 13 days ago	Sprint 2: Main ...						
② Create: "2.Objetivos" documentation	documentation						
#11 by aguadotzn was closed 23 days ago	Sprint 1: Invest...						
② Create: "1.Introducción" documentation	documentation						
#10 by aguadotzn was closed 23 days ago	Sprint 1: Invest...						
② Research about Air Quality (Environmental legislation) investigation	investigation						
#9 by aguadotzn was closed 13 days ago	Sprint 1: Invest...						
② Research about Air Quality (Science) investigation	investigation						
#8 by aguadotzn was closed 13 days ago	Sprint 1: Invest...						
② Send emails to recap information documentation investigation	documentation investigation						
#7 by aguadotzn was closed 23 days ago	Sprint 1: Invest...						
② Analyze data data investigation	data investigation						
#6 by aguadotzn was closed 13 days ago	Sprint 1: Invest...						
② Discuss proposal with Carlos (Indra) investigation	investigation						
#5 by aguadotzn was closed 13 days ago	Sprint 1: Invest...						
② Talk with Carlos (academic purposes) university Stuff	university Stuff						
#4 by aguadotzn was closed 23 days ago	Sprint 1: Invest...						

Figura A.1: Detalle *Product Backlog*.

Al inicio de cada Sprint se extraen una serie de tareas de los elementos del *Product Backlog* que conforman el *Sprint Backlog*, las tareas son elegidas por el equipo. Estas tareas deben ser completadas durante el ciclo y no es posible añadir tareas nuevas a este *Sprint* durante el tiempo que se está realizando el mismo.

- *Product Backlog*: en un entorno real, consistiría en un listado de historias de usuario, obtenidas con el cliente, que se irán incorporando al producto a partir de incrementos sucesivos. En el proyecto descrito aquí éste se ve reflejado en las tareas (*issues* en github) que se van creando a medida que son necesarias. Estás tareas se asignan después a cada sprint.
- *Sprint*: una de las bases de los proyectos ágiles es el desarrollo mediante las iteraciones incrementales. En Scrum a cada iteración se le denomina *Sprint*. En el caso concreto de este trabajo, excepto el último, todos tienen una duración de quince días.

A continuación se describen las reuniones que se realizan en el marco de desarrollo de trabajo *Scrum*.

- *Daily Meeting*: se realiza una reunión breve diaria en la que cada miembro del equipo cuenta el estado en el que se encuentra la parte del proyecto que está realizando. En el caso del proyecto que se describe en esta documentación deberá ser una reflexión personal.
- *Planning Meeting*: se realiza al inicio de cada *Sprint* con el objetivo de escoger las tareas necesarias a realizar. Recodemos que se deben escoger del *Product Backlog*.
- *Review Meeting*: esta reunión se realiza con el cliente al final de cada *Sprint*, suele tener una duración máxima de cuatro horas y se revisa las historias de usuarios que han finalizado y las que no. Además se suele realizar una pequeña demo.
- *Retrospective Meeting*: esta reunión también se realizala finalizar el *Sprint*. Todos los miembros del equipo expresan sus opiniones sobre la iteracción finaliza. El objetivo de esta reunión es realizar una mejora continua del proceso.

A continuación se describen los diferentes *sprints* que se han realizado de manera detallada así como las planificaciones inicial y final.

Planificación inicial

Incluir aqui diagrama gannttt INICIAL

Mayo/Junio 2019

El trabajo fin de máster fue elegido a finales de marzo de 2019. Debido a la intesa carga de trabajo del máster y a que la actividad del autor no se limita únicamente al ámbito académico sino que también desempeña un trabajo a tiempo completo durante la semana, no se pudo empezar a desarrollar una actividad completa hasta el mes de junio. En la figura A.2 se puede ver el *board* al inicio.

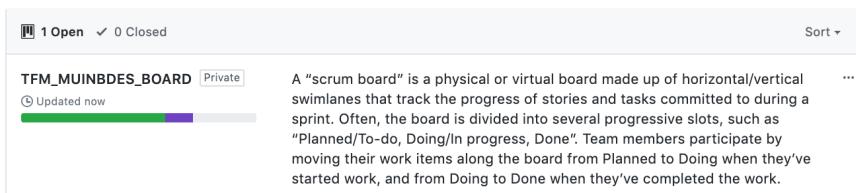


Figura A.2: Detalle *Scrum board*.

Sprint 0: 15/06/2019 - 01/07/2019

- * Duración: 15 días
- * Descripción Sprint: Este primer *Sprint* no se puede considerar como tal, de ahí el número cero, ya que no estaba al cien por cien en el trabajo fin de máster. Mi prioridad eran las asignaturas ya que sin ellas no podría desarrollarlo. Durante estos primeros quince días me dediqué a crear tareas y por tanto a llenar el *backlog*. Si bien es cierto que la mayoría eran tareas como cosas de la documentación, cosas muy obvias sobre el desarrollo o tema académico. Durante la continuación del desarrollo se iba llenando el backlog con nuevas tareas que me iban surgiendo. De la misma manera también mucho tiempo dedicado a leer durante este *Sprint*.

Sprint 1: 01/07/2019- 15/07/2019 (ver figura A.3)

- * Duración: 15 días
- * Descripción Sprint: En este primer *Sprint* efectivo se comienza a trabajar en hacer un criba sobre todo lo investigado, además se comienza también a crear la documentación, repositorios y a hacer un análisis preliminar de los datos con los que vamos a trabajar.

1. Planning meeting

Reseña: La primera reunión se fijan las tareas que se van a realizar este *Sprint* y se concuerda con el tutor académico lo que tenemos en mente para comenzar con el desarrollo.

2. Sprint planning

- En esta reunión se establecen las tareas (*user stories*) a desarrollar durante el *Sprint*.
 - Crear repositorio.
 - Investigar tecnologías principales.
 - Analizar datos.
 - Discutir propuesta con tutor académico.
 - Discutir propuesta con tutor empresa.
 - Terminar investigación sobre legislación.
 - Terminar investigación científica sobre calidad del aire.
 - Documentación: Introducción.
 - Documentación: Objetivos.

3. Retrospective meeting

- ¿Qué ha ido bien?
 - Toda la parte de investigación tanto sobre legislación como científica.
 - Reuniones muy productivas con los tutores para aclarar diversas dudas de índole diferente.
 - Comienzo con la memoria de manera constante y paralela a las demás partes.
- ¿Qué dificultades hemos encontrado?
 - Debido a la enorme cantidad de información teórica que he recabado me va a ser complicado volcarla toda rápida
 - Existen numerosísimas opciones en cuanto a tecnologías. La oferta en el mercado es muy amplia y ello me crea incertidumbre al respecto.

Sprint 1: Investigation

⚠ Past due by 16 days 100% complete

Continue with the investigation about climate change, data, air quality, laws, academic stuff...

	0 Open ✓ 11 Closed
≡	ⓘ Research about Air Quality (Environmental legislation) investigation #9 by aguadotzn was closed 5 days ago
≡	ⓘ Research about Air Quality (Science) investigation #8 by aguadotzn was closed 5 days ago
≡	ⓘ Discuss proposal with Carlos (Indra) investigation #5 by aguadotzn was closed 5 days ago
≡	ⓘ Analyze data data investigation #6 by aguadotzn was closed 5 days ago
≡	ⓘ Create: "2.Objetivos" documentation #11 by aguadotzn was closed 15 days ago
≡	ⓘ Create main documentation documentation #3 by aguadotzn was closed 15 days ago
≡	ⓘ Send emails to recap information documentation investigation #7 by aguadotzn was closed 15 days ago
≡	ⓘ Talk with Carlos (academic purposes) university Stuff #4 by aguadotzn was closed 15 days ago
≡	ⓘ Create: "1.introducción" documentation #10 by aguadotzn was closed 15 days ago
≡	ⓘ Investigate main technologies investigation #2 by aguadotzn was closed 15 days ago
≡	ⓘ Create repository devOps #1 by aguadotzn was closed 15 days ago

Figura A.3: Detalle Sprint 1.

Sprint 2: 15/07/2019- 31/07/2019(ver figura A.4)

* Duración: 15 días

- * Descripción Sprint: En este *Sprint* se ha enfocado más a comprender, analizar e interpretar toda la información de la que se disponía. Ha sido un sprint de escribir mucho, asimilación y análisis antes de volver a la parte técnica. A veces es necesario tener claro todos los conceptos para saber como avanzar.

1. *Planning meeting*

Reseña: En esta reunión se fijan las tareas que se van a realizar este *Sprint*. Además se añaden algunas nuevas al backlog como las referidas a la parte técnica que se realizarán en scripts sucesivos y que vienen derivadas de las necesidades de éste *Sprint*.

2. *Sprint planning*

- En esta reunión se establecen las tareas (*user stories*) a desarrollar durante el *Sprint*.

- Creación de scripts y arquitectura del TFM.
- Preprocesamiento y limpieza de datos.
- Documentación: conceptos teóricos (Big Data).
- Documentación: conceptos teóricos (Legislación).
- Documentación: conceptos teóricos (Método análisis).
- Documentación: conceptos teóricos (Análisis científico).

3. *Retrospective meeting*

- ¿Qué ha ido bien?
 - Con este sprint se ha comprendido una de las partes más importantes del trabajo: la parte científica.
 - Toda la parte de preprocesamiento de datos.
- ¿Qué dificultades hemos encontrado?
 - Debido a la extensión de los conceptos teóricos no he podido progresar todo lo que me hubiera gustado en la parte técnica. Será necesario refactorizar código en *sprints* sucesivos.

Sprint 2: Main development 1/2 (preprocess data)

Due by July 31, 2019 100% complete

After the investigation I need to process the data in order to extract only the information that is useful to the thesis main objective.

	0 Open ✓ 6 Closed
Preprocess data	
Create scripts.	
Create "3.Conceptos teóricos" (Part IV: theoretical introduction to technical concepts) documentation	
Create "3.Conceptos teóricos" (Part II: Scientific analysis) documentation	
Create "3.Conceptos teóricos" (Part III: Legislative analysis) documentation	
Create "3.Conceptos teóricos" (Part I: Big Data) documentation	

Figura A.4: Detalle Sprint 2.

Sprint 3: 31/07/2019 - 15/08/2019 (ver figura A.5)

- * Duración: 15 días
- * Descripción Sprint: Este tercer *Sprint* se ha dirigido más a la parte técnica. Tanto seguir y terminar tareas del *sprint* anterior como en la parte de visualización.

1. *Planning meeting*

Reseña: En esta reunión he escogido las tareas a realizar en este sprint.

2. *Sprint planning*

- En esta reunión se establecen las tareas (*user stories*) a desarrollar durante el *Sprint*.
 - Visualización de datos en *Python*.
 - Visualización de datos en *R*.
 - Refactorización de código.
 - Documentación: aspectos relevantes del desarrollo.
 - Documentación: técnicas y herramientas.
 - Documentación: trabajos relacionados.
 - Enviar copia a tutores de los puntos de la memoria desarrollados en el *sprint* anterior.

3. *Retrospective meeting*

- ¿Qué ha ido bien?
 - Este *sprint* se ha enfocado en trabajar las visualizaciones tanto en *Python* como en *R*.
 - Se ha avanzado mucho en la memoria y además se ha enviado una primera copia a los tutores para su revisión.
- ¿Qué dificultades hemos encontrado?
 - Ninguna dificultad destacada.

Sprint 3: Main development 2/2 (visualize data)

⚠ Past due by 3 days 100% complete

Once we have all our data already preprocessed we have to focus in visualize it. Investigation about the best solutions in data visualization.

	0 Open	✓ 7 Closed
■ ⌚ Data visualization (python): exploring frameworks	data-visualization	investigation
#17 by aguadotzn was closed 1 minute ago		
■ ⌚ Data visualization (R): exploring frameworks	data-visualization	investigation
#18 by aguadotzn was closed 1 minute ago		
■ ⌚ Send first draft copy to Carlos.	documentation	meeting
#27 by aguadotzn was closed 5 days ago		
■ ⌚ Create "5.Aspectos relevantes del desarrollo"	documentation	
#21 by aguadotzn was closed 8 days ago		
■ ⌚ Create "6.Trabajos relacionados"	documentation	
#22 by aguadotzn was closed 8 days ago		
■ ⌚ Create "4.Técnicas y herramientas"	documentation	
#20 by aguadotzn was closed 8 days ago		
■ ⌚ Refactor code.	data	data-extraction
#28 by aguadotzn was closed 11 days ago		data-processing

Figura A.5: Detalle Sprint 3.

Sprint 4: 15/08/2019 - 31/08/2019 (ver figura ??)

- * Duración: 15 días
- * Descripción Sprint: Este cuarto *Sprint* se ha dirigido más a la parte de la visualización final aunque sin olvidar la documentación.

1. *Planning meeting*

Reseña: En esta reunión he escogido las tareas a realizar en este *sprint*.

2. *Sprint planning*

- En esta reunión se establecen la tareas (*user stories*) a desarrollar durante el *Sprint*.

- Visualización de datos con PowerBI.

- Investigación de uso de la herramienta software *tinybird*.
- Discusión de resultados.
- Documentación: plan de proyecto software.
- Documentación: entorno experimental.
- Documentación: resultados.
- Enviar segunda copia de la memoria a los tutores

3. *Retrospective meeting*

- ¿Qué ha ido bien?
 - Este *sprint* se ha dirigido a trabajar en la parte de la visualización final.
 - Se ha avanzado en la memoria respecto de las partes finales de la misma.
 - Al introducirse una nueva herramienta (*tinybird*) se ha estado investigando su funcionamiento.
- ¿Qué dificultades hemos encontrado?
 - La parte de la visualización final se ha comido demasiado tiempo por lo que todas las tareas de documentación no han podido finalizarse tal y como se esperaba.

Sprint 5: 31/08/2019- 05/09/2019 (ver figura ??)

Planificación final

Incluir aqui diagrama gannttt FINAL

A.3. Viabilidad legal

La viabilidad legal se centra principalmente en el estudio de las licencias software utilizadas. Realizaremos una tabla resumen sobre las licencias. Indicar aquí que al tratarse de un estudio analítico se ha intentado seguir e utilizar software de carácter *open source*.

Dependencias	Licencia
<i>Github pages</i>	MIT
<i>R</i>	GNU
<i>Python</i>	PSFL
<i>Vega</i>	BSD-3 Clause

Tabla A.1: Tabla resumen-licencias.

Una licencia software es un contrato entre el autor o titular de los derechos de explotación o distribución y el usuario consumidor, usuario profesional o empresa, para la utilización del software cumpliendo una serie de términos y condiciones establecidas dentro de sus cláusulas. Todo el software que empleo es su amplia mayoría es libre ya que la mayoría de código viene o bien de desarrolladores *amateur* o es libre desde su creación. Emplean por tanto las licencias descritas en la tabla anterior por lo que diferencia entre ellas varían en términos como el nombramiento del autor o la garantía , son licencia comunes en software libre [15] , además github provee una página para ayudarte en la elección de tu licencia [21]. La menos permisiva de entre las descritas es *MIT* [25] y la más es la licencia *Apache-2.0* [14].

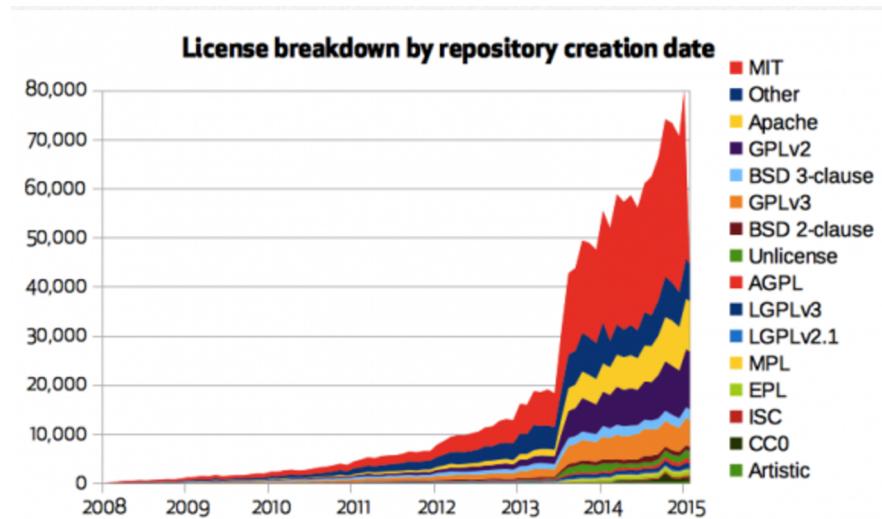


Figura A.6: Gráfico licencias Open Source in GitHub. Fuente: <https://cartograf.net>.

Para el proyecto de manera general (está alojado en github de forma libre) he elegido la licencia [MIT](#).

Para la documentación he elegido una licencia Creative commons, en concreto se ha elegido la *Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)*.

Reconocimiento-NoComercial 4.0 Internacional



Figura A.7: Licencia Creative Commons.

A.4. Links Importantes