Madrid, isla de calor
Madrid, heat island

Trabajo de Fin de Grado
Curso 2022–2023

Autor
José Francisco García Ruiz
Gonzalo Meneses Vicente

Directora
Sonia Estévez Martín

Grado en Ingeniería Informática e Ingeniería del Software
Facultad de Informática
Universidad Complutense de Madrid

Madrid, isla de calor
Madrid, heat island

Trabajo de Fin de Grado en Ingeniería Informática e Ingeniería del Software

Autor
José Francisco García Ruiz
Gonzalo Meneses Vicente

Directora
Sonia Estévez Martín

Convocatoria: *Febrero 2023*

Grado en Ingeniería Informática e Ingeniería del Software
Facultad de Informática
Universidad Complutense de Madrid

23 de FEBRERO de 2022

Dedicatoria

A Sergio, Carlos, Mariángeles y toda mi familia y amigos que sin su apoyo diario, amor incondicional y energía no hubiese sido posible ninguno de mis logros.
Gonzalo.

A mi pareja, amigos y toda mi familia, por que nunca es tarde si la dicha es buena.
José.

Agradecimientos

«Queremos agradecer a nuestras familias todo el apoyo, fuerza y cariño que nos han dado durante nuestras vidas, durante este trabajo y siempre que lo hemos necesitado. También todos los profesores que hemos tenido, tanto en la universidad como fuera de ella, por haber fomentado, aunque algunos más que otros, el desarrollo de nuestra curiosidad. En particular a Sonia por habernos guiado, aconsejado y motivado en este punto y final de la carrera. También queremos agradecer a aquellos que nos han formado como profesionales y como personas, realizando un trabajo que nunca se podrá valorar lo suficiente. Y por último agradecer a todos los grandes amigos que hemos hecho fuera y dentro de la Facultad, por los buenos momentos juntos, las alegrías compartidas, la buena energía transmitida y por tendernos una mano cuando la hemos necesitado durante estos años.»

Resumen

Madrid, isla de calor

Madrid como tantas otras ciudades, es una *" isla de calor "*. El calentamiento del asfalto y de los edificios hacen que la temperatura suba en unas zonas más que en otras, algo que podemos comprobar cuando nos movemos entre diferentes distritos de la ciudad.

En este trabajo se pretende estudiar las relaciones entre los distintos contaminantes, la meteorología y las zonas verdes e hidrográficas, empleando datos históricos recogidos por las estaciones que se encuentran por la ciudad de Madrid. Aplicando algoritmos de Machine Learning podremos determinar si esta correlación afecta en las temperaturas de las distintas zonas de la ciudad.

Palabras clave

Contaminación, meteorología, zonas verdes, zonas hidrográficas, datos abiertos, aprendizaje automático, isla de calor.

Abstract

Madrid, heat island

Madrid, like so many other cities, is a ”heat island”. The heating of the asphalt and buildings makes the temperature rise in some areas more than in others, something that we can see when we move between different districts of the city.

The aim of this work is to study the relationships between different pollutants, meteorology and green areas and hydrographic zones, using historical data collected by the stations located in the city of Madrid. By applying Machine Learning algorithms we will be able to determine if this correlation affects the temperatures of the different areas of the city.

Keywords

Pollution, meteorology, green areas, hydrographic areas, open data, machine learning, heat island.

Índice

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La temperatura media en las grandes ciudades se va superando cada año, estableciendo nuevos récords. Sin embargo, en pequeñas urbes y pueblos no tan alejados de dichas áreas, no pasa este fenómeno. ¿A qué se debe? La respuesta está en las *Isla de Calor Urbanas* (ICU en adelante).

Este problema es consecuencia de una gran variedad de factores. Empezando por la gran cantidad de edificaciones y materiales de construcción, que absorben calor y lo irradian lentamente. Además de la contaminación generada por vehículos y sistemas de climatización. También, la escasez de zonas verdes y fluviales aumenta este fenómeno. Esto, sumado al cambio climático, aumenta las temperaturas, que implica un crecimiento en el consumo energético de la población. Pero no solo afecta en temas económicos, sino que también impacta sobre la salud de los ciudadanos, agravando patologías, empeorando la salud mental, etc. Sierra (2021).

Madrid, que cuenta con 3 305 408 personas (INE 2021), es la urbe más habitada de España y la segunda de la Unión Europea[1] . Además, la M-30, A-2 y M-40 son carreteras con gran afluencia de tráfico. Todo esto la sitúa dentro de las ciudades con peor calidad del aire de España[2], lo que repercute en que el efecto ICU se intensifique.

Gracias al Portal de Datos Abiertos[3] del Ayuntamiento de Madrid podemos acceder a un amplio *dataset* que almacena información sobre el municipio. Para este TFG usaremos datos de las distintas estaciones meteorológicas y de calidad del aire de los distritos que componen la ciudad. Después de un tratamiento de los datos, vamos a analizar cómo se distribuye el efecto ICU en Madrid sirviéndonos de algoritmos de *Machine Learning* (ML en adelante), como el *KMeans*, y de representaciones en gráficos como dendrogramas, y formas más visuales como mapas.

[1] Enlace Datos Población Unión Europea
[2] Articulo Ciudades con peor calidad del aire
[3] Enlace al Portal de Datos Abiertos del Ayuntamiento de Madrid

1.1. Motivación

La principal idea que nos llevó a realizar este trabajo fue la enorme diferencia de las temperaturas entre el centro de Madrid y el extrarradio y las ganas de explicar este suceso, ya que en cuestión de kilómetros la temperatura cambia de media unos 3ºC. Pero lo más impactante fue que en días de temperaturas extremas, como puede llegar a pasar en invierno, la diferencia llega a superar los 15ºC, como se puede ver en la Figura 1.1.

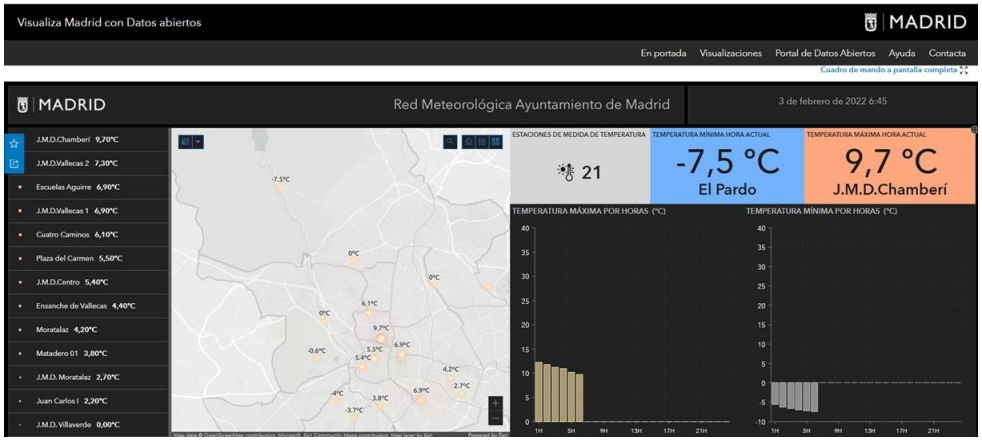


Figura 1.1: Diferencia de 17.2ºC entre Madrid centro y el Pardo

Además, el hecho de que se trate de la ciudad en la que residimos y que este suceso llamado *ICU* nos afecte directamente a nosotros y nuestra calidad de vida, nos impulsó a querer analizar los porqués de estas diferencias.

Aparte de lo mencionado, también tenemos un interés personal. A lo largo de nuestra formación universitaria en la facultad hemos desarrollado un interés sobre el *Big Data* y el *ML*, lo que nos incentivó a realizar este trabajo.

1.2. Objetivos

El principal objetivo del trabajo es la aplicación de algoritmos de *ML* para detectar que zonas se comportan de una forma semejante en cuanto a variables meteorológicas y urbanísticas de Madrid y las razones que lo producen. Los objetivos se podrían resumir en:

- O1. Realizar técnicas de limpiado, formateado y filtrado de datos a gran escala, recogidos del Portal de Datos Abiertos del Ayuntamiento de Madrid, para la obtención de unos datos homogéneos analizables.
- O2. Aplicar algoritmos del mundo del *Data Science* para poder realizar distintas clasificaciones de las estaciones meteorológicas.
- O3. Uso de herramientas de visualización de datos para ilustrar estas agrupaciones para una mayor facilidad a la hora de interpretar los resultados.

- O4. Obtener conclusiones de las clasificaciones gracias a la visualización y proponer soluciones.

La obtención de datos, exploración y procesado de los mismos, elección de algoritmos, visualización y análisis, son fundamentales dentro del *Data Science*.

Otro de los objetivos, aunque personal, es aprender a programar en un lenguaje nuevo para ambos, como es *R*, así como las herramientas necesarias para la limpieza, clasificación y visualización mediante el uso de librerías en *RStudio* y comprender el contexto de las diferencias de temperaturas en nuestra ciudad de residencia, Madrid.

1.3. Plan de trabajo

Gracias a los conocimientos que hemos adquirido en asignaturas impartidas y en los proyectos realizados durante la realización de los diferentes Grados, especialmente en las de Ingeniería del Software, hemos sido capaces de hacer una planificación que hemos considerado bastante útil y resolutiva. Siendo un equipo pequeño, dos alumnos y una tutora, y no hayamos podido aplicar los métodos de desarrollo ágil tal y como se nos han sido enseñados, hemos intentado adaptarlo al tamaño del equipo.

Nuestro trabajo se ha dividido en dos principales ramas, un borrador de la memoria y el código.

- **Borrador de la memoria**: puesto que desde un principio se nos ha avisado de la importancia de la memoria final, vimos conveniente generar un documento en el que ir redactando toda la información que recabábamos, explicaciones de los formatos de datos y del tratamiento del mismo, explicaciones del código y observaciones y enlaces de interés. Este documento ha sido muy útil a la hora de redactar la memoria, ya que nos ha facilitado el acceso a la información que hemos recabado, así como los pasos que hemos seguido para llegar hasta finalizar el trabajo.
- **Código**: como todo desarrollo software siempre es necesario el uso de un repositorio que controle las versiones que se van sacando a lo largo del proceso. Tanto es así que decidimos crear uno en la plataforma **GitHub**[4] donde hemos estado subiendo el código perteneciente a los distintos objetivos marcados.

1.3.1. Lenguaje de desarrollo: R

El enfoque que le queríamos dar al trabajo era más estadístico con partes de análisis y visualización sobre grandes conjuntos de datos por lo que en la fase de investigación contemplamos distintos lenguajes en los que programar, pero como queríamos recalcar en una visión estadística decidimos emplear *R*. Además, puesto que la aplicación de algoritmos de *ML* iban a ser muy relevantes para la clasificación de las estaciones meteorológicas, nos reafirmamos en nuestra decisión. Esto se debe a que hay una gran cantidad de librerías, e incluso en R base, en las que se incluyen funciones que implementan dichos algoritmos.

[4]Enlace al repositorio, https://github.com/Gmene00/TFG-Madrid-Isla-de-calor

1.3.2. Entorno de desarrollo: RStudio

Una vez elegido el lenguaje, escoger un entorno de desarrollo es mucho más fácil. Había muchas opciones, desde *VisualStudio*, *Geany*, *RKWard*, etc. Aparte de que la mayoría de los tutoriales online usaban RStudio, decidimos guiarnos por la recomendación de nuestra tutora, Sonia, por lo que este trabajo ha sido completamente desarrollado con el *IDE RStudio*. Esto ha sido de gran utilidad ya que nos muestra todos los objetos del *workspace* así como los comandos ejecutados ya sean en *scripts* o en la consola que proporciona, así como un visor con los paquetes disponibles e instalados. Esta herramienta cuenta con un visor de gráficos, que, junto a la extensa cantidad de librerías disponibles, hacen que la visualización sea intuitiva, sencilla y rápida.

Estas son las principales librerías empleadas en el trabajo:

- Paquete *Tidyverse*, es una colección de paquetes y librerías diseñadas para el *Data Science*, que nos permite manipular los datos para una mejor interpretación tanto en tablas como en gráficos. De esta colección hemos usado los siguientes paquetes:
  - Librería *ggplot2*: permite crear gráficos declarativamente. Basta con proporcionarle los datos y las variables de estética y genera fácilmente la figura. Como complemento, está *ggpubr* que aumenta las posibilidades para mostrar los datos de una forma más elegante.
  - Librería *dplyr*: permite manipular los datos de una forma sencilla, sobre todo gracias a la función *select()*, que es capaz de devolver subconjuntos de otros más grandes en una sola sentencia.
  - Librería *tidyr*: al limpiar datos resultan útiles muchas de sus funciones que eliminan los *valores nulos*, así como al formatear ofrece funciones para dividir celdas, remodelar *dataframes*, etc.
- Librería *cluster*: Como lo define la propia librería, proporciona métodos para el análisis de *clusters*.
- Librería *factoextra*: contiene funciones que aplican algoritmos de *ML* y facilitan la visualización, como el *diagrama del codo*.
- Librería *osmdata*: permite descargar e importar información de OpenStreetMap como objetos sp o sf. Nos permite realizar los mapas.
- Librería *sp*: proporciona clases y métodos para tratar con datos espaciales.
- Librería *ggmap*: añade a los gráficos ya conocidos, los extraídos con elaboración propia, una capa cartográfica adicional.

1.3.3. Método de desarrollo: Cascada

Para seguir con un correcto desarrollo del *TFG*, vemos conveniente especificar como hemos estado estructurando, planificando y controlando el proyecto a lo largo

del tiempo. Para ello hemos intentado seguir los pasos de las metodologías de desarrollo ágiles, aunque estas están recomendadas para grupos de entre cinco y nueve personas, la adaptamos a nuestras necesidades para poder estructurarnos de una manera adecuada que a continuación explicamos.

Los objetivos del trabajo son dependientes unos de otros, es decir, no podemos aplicar los algoritmos de *ML* sin antes haber limpiado y formateado los datos, así como no se pueden visualizar los resultados en forma de gráfica ni en los mapas sin antes haber aplicado los propios algoritmos. Esto nos ha hecho elegir la metodología *Waterfall* o en *Cascada*, que se caracteriza por dividir el proyecto en funciones diferenciadas, lo que para nosotros han sido los objetivos.

Esta metodología requiere que antes de pasar a la siguiente función u objetivo, el proyecto debe ser revisado y en caso de encontrar fallos ser corregido. Es aquí donde entra la tutora del *TFG*, Sonia, con la que hemos concertado reuniones aproximadamente cada dos semanas, de manera presencial en su mayoría. Donde éramos orientados en cada etapa de la *cascada* y revisaba que los objetivos marcados estaban siendo realizados.

Otra de las ventajas de haber seguido este modelo es que, al estar los objetivos bien definidos desde bien temprano en el desarrollo, la planificación ha sido sencilla.

Además, para una mayor organización y control sobre fechas, así como reparto de tareas, hemos recurrido a un diagrama de *Gantt* que nos ha permitido tener una mayor consciencia sobre el trabajo realizado hasta cada reunión, al igual de los objetivos pendientes. A modo de resumen de la planificación, la Figura 7.3 muestra como ha sido el proceso.

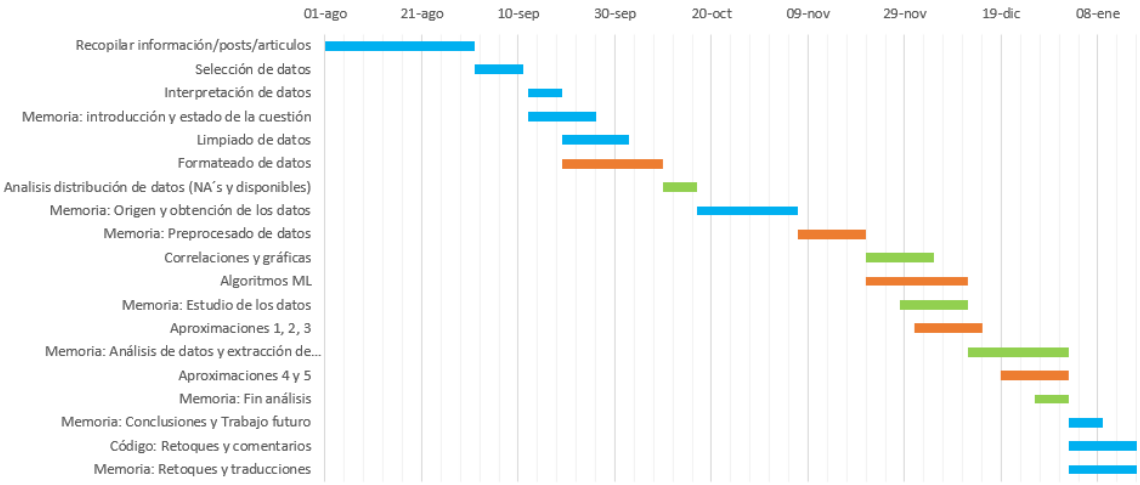


Figura 1.2: Trabajo; azul-*Ambos*, verde-*José*, naranja-*Gonzalo*

1.4. Contenido de la memoria

La presente memoria consta de los siguientes apartados:

1. Introducción: en este capítulo se plantean los problemas, objetivos y mecanismos para alcanzarlos.

2. Estado de la cuestión: durante este capítulo se presentarán las diferentes zonas de Madrid y su firma de calor, contextualizando y clasificando cada una de dichas zonas, aparte del estado actual de diversos trabajos que abordan su estudio. También se presentarán las principales diferencias de este proyecto con otros proyectos relacionados.

3. Origen y obtención de los datos: en este apartado se presentará la fuente de donde se han extraído los datos, y la metodología utilizada para su obtención. Además, también se explica la descripción de estos en su forma original.

4. Preprocesado de los datos: donde se explicarán los métodos utilizados para su tratamiento y limpieza, para poder lograr un formato adecuado para su estudio, y de este modo facilitar la extracción de información. También se muestra el resultado obtenido tras la transformación.

5. Estudio de los datos: se muestran las distintas técnicas y métodos para el tratamiento de los valores nulos y valores inesperados o outliers, con el objetivo de no perder demasiada información relevante y eliminar aquella que pueda introducir ruido en los resultados.

6. Análisis de datos y extracción de información: en este apartado exponemos los algoritmos que se han utilizado para la creación y selección de modelos y las técnicas que se pueden utilizar para mejorar las predicciones de estos. También, se presentará la comparación entre los distintos modelos.

7. Conclusiones y trabajos futuro: en esta sección se enumeran las conclusiones alcanzadas tras los resultados obtenidos durante el apartado anterior. Además, se describen posibles trabajos futuros que se podrían realizar a raíz de este proyecto.

Capítulo 2

Estado de la Cuestión

En este capítulo reunimos información y referencias bibliográficas que nos han servido de utilidad para poner en contexto nuestro trabajo. Entre otros serian: el número de habitantes, densidad de población, superficie, etc. Hay que añadir otros aspectos relevantes como las zonas verdes, la proximidad a carreteras con tráfico y el efecto de los contaminantes producidos por los vehículos y la maquinaria climatizadoras. Así como las magnitudes meteorológicas que son más relevantes al analizar el efecto *ICU*.

Para poder estudiarlo desde una mejor perspectiva hemos seleccionado una serie de estaciones meteorológicas y de calidad del aire representativas de la ciudad. El Ayuntamiento de Madrid nos ofrece un conjunto de datos abiertos con los valores recogidos con las magnitudes de nuestro interés. Todo este proceso está explicado en el punto de obtención de los datos de forma más específica.

Resulta necesario conocer la ubicación de las estaciones, ya que no es lo mismo una estación meteorológica o de calidad del aire ubicada en el parque *Juan Carlos I*, que otra estación ubicada en *Cuatro Caminos*. La primera destaca por estar rodeada de zonas verdes, con la presencia de una gran zona fluvial. Mientras que la segunda se encuentra en uno de los barrios más céntricos de la ciudad donde encontramos un elevado número de edificios, colindante con el *Paseo de la Castellana* arteria central de Madrid que concentra un tránsito alto de vehículos. Esto hace que haya una diferencia de temperaturas a lo largo del año, pero más notoria en verano e invierno.

2.1. Estaciones

Anteriormente hemos hablado sobre la relevancia de la localización de las estaciones, vamos a poner en contexto todas aquellas que nos han servido en nuestro estudio. Para ello contamos con el mapa de la Figura 2.1. Este ubica todas las estaciones tanto meteorológicas como de calidad del aire, de las que hablaremos en los siguientes puntos.

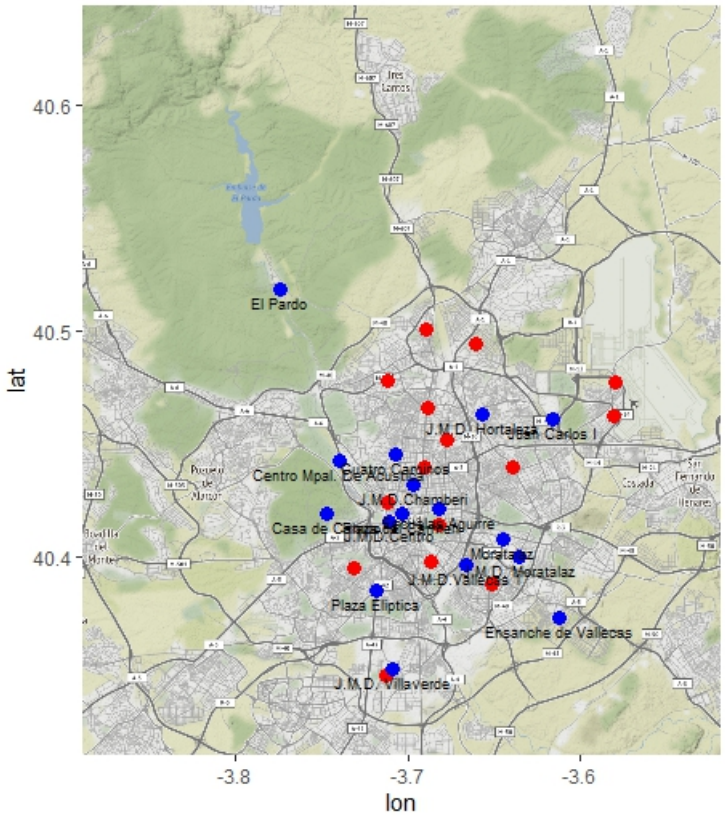


Figura 2.1: Azul son las estaciones meteorológicas y Rojo estaciones de calidad del aire

2.1.1. Estaciones de meteorología

De las 26 estaciones con las que cuenta el **Sistema Integral de Calidad del Aire**[1], y debido a que no todas las estaciones tienen los mismos sensores meteorológicos -como se puede observar en el portal-, decidimos escoger las estaciones que muestra la Figura 2.2, por su ubicación y medidas recogidas:

Toda la información a continuación mencionada es extraída de **Distrito en Cifras del ayuntamiento de Madrid**[2]

- Moncloa-Aravaca: es uno de los distritos más extensos de Madrid ya que cuenta con Casa de Campo en sus límites, siendo el parque público más grande del municipio y cuenta con 3555,6 *ha* de zonas verdes y un gran lago, influyendo en las temperaturas y la calidad del aire. Esto le proporciona al distrito una superficie de 46,53 km^2, y con una población de 116.903 habitantes posee una densidad de 2512,36 hab/km^2. Encontramos dos estaciones en su interior:
  - Estación del Centro Municipal De Acústica: en plena M-30, una de las carreteras con mayor congestión, en un núcleo urbano donde podemos encontrar numerosas facultades y edificios municipales. Esto le asegura una densidad de habitantes y vehículos elevada.

[1]Enlace Sistema Integral de Calidad del Aire
[2]Enlace Distrito en cifras del Ayuntamiento de Madrid

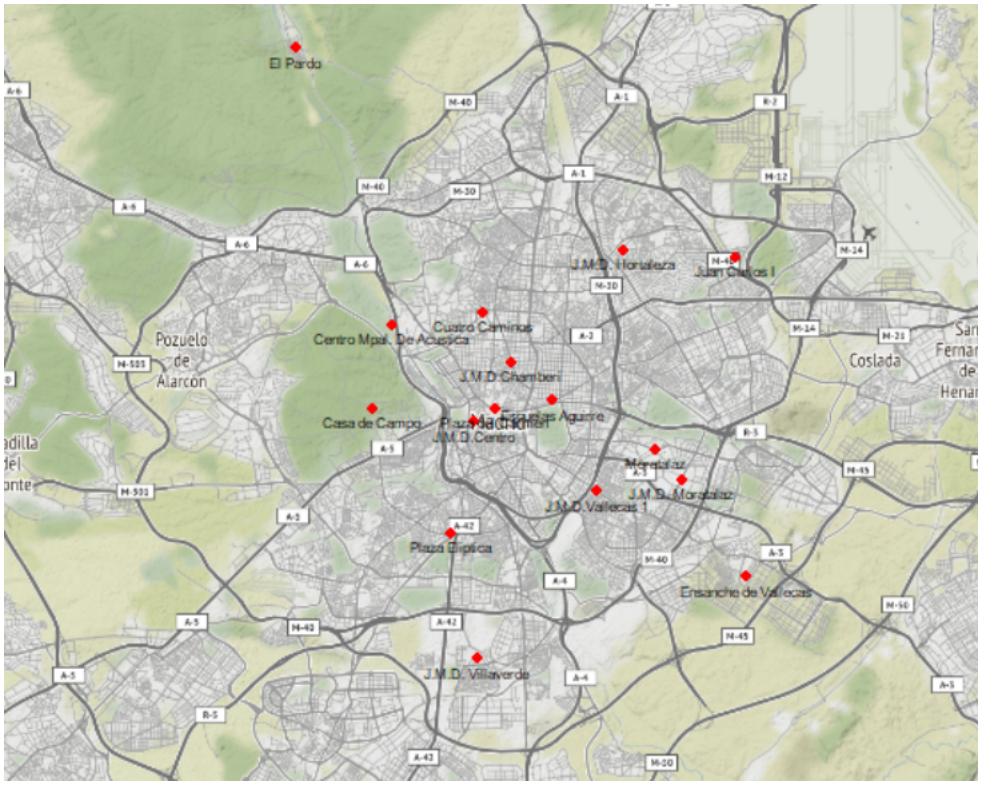


Figura 2.2: Ubicación de las estaciones meteorológicas

- Estación de Casa de Campo: localizada en el interior del parque y próxima al lago, se presupone como una de las zonas con mejor calidad del aire.

- Centro: es la zona más antigua de Madrid conteniendo numerosos puntos de interés tanto para habitantes y turistas. Esto hace que sea el distrito más transitado. Con una superficie de 5,2373 km^2 y 140.473 habitantes, se coloca como uno con mayor densidad de población. Esto hizo que el Ayuntamiento tomase la decisión de nombrarlo zona de bajas emisiones para paliar la elevada contaminación. Como consecuencia el tránsito de vehículos ha bajado considerablemente en los últimos años.

  - Estación de la Junta Municipal del Distrito Centro: la J.M.D. se encuentra a escasos 200 metros del Palacio Real, así como del Teatro Real, cerca de sus parques y jardines.
  - Estación de la Plaza del Carmen: esta plaza se encuentra entre los sitios más concurridos de la ciudad, Gran Vía y Sol. Al ser calles peatonales principalmente están destinadas a comercios, por lo que la presencia de zonas verdes es más reducida pero también se ve reducido el número de vehículos.

- Moratalaz: tiene una población de 95.036 personas y una superficie 6,34 km^2, lo que le da una densidad de población de 14.989 hab./km^2. Está localizado al este del municipio de Madrid, y de su superficie, 1,3192 km^2 (un 20,9 %) se

corresponden a zonas verdes. Está delimitado por cuatro autopistas: M-30 al oeste; M-40 al este; A-3 al sur y la M-23 (eje O'Donnell) al norte.

  - Estación de la J.M.D. Moratalaz: es una zona puramente residencial, con numerosas avenidas muy concurridas tanto por vehículos como personas. Este es el caso de la Calle de la Fuente Carrantona donde se ubica la estación.
  - Estación de Moratalaz: al contrario que la anterior, en las inmediaciones de esta predominan las zonas verdes. Situada en el Parque de la Cuña Verde de O´Donnell, y separada del Cementerio Municipal Nuestra Señora de la Almudena por la M-23. Esto hace que también se presuponga como otra de las zonas con mejor calidad del aire en Madrid.

- Estación de Ensanche de Vallecas, en Villa de Vallecas, tiene una población de 107.649 habitantes y con una extensión de 51,47 km^2 supone una densidad de 2091,6 hab/km^2. Esta gran extensión tiene como beneficio sus grandes zonas verdes, que gozan de casi 750.000 m^2, no obstante, también tiene un gran solar al sur el cual carece prácticamente de masa arbórea. Está delimitado en su gran mayoría por carreteras como la M-50 al sur, A-3 al este, al norte la M-40. Aunque eso sí, limita al oeste por el río Manzanares.

- Estación de la J.M.D.Vallecas, situada en Puente de Vallecas. Es uno de los distritos más poblados de la capital, con 240.867 habitantes. Cuenta con grandes zonas verdes como el Parque Forestal de Entrevías y el Cerro del Tío Pío. Aun así, dichas zonas no afectan demasiado a la estación ya que esta se encuentra en pleno núcleo urbano, en la Avenida de la Albufera y a escasos 100 metros de la M-30.

- Estación de la J.M.D. Hortaleza, situada en Hortaleza. Con 27,4198 km^2, es otro de los grandes distritos. Pero con sus 193.264 habitantes, hace que su densidad, 6581,46 hab/km^2 no sea muy elevada. Además, cuenta con grandes zonas verdes como el Parque Juan Pablo II, el Parque Forestal de Valdebebas y el Pinar del Rey, en el que se ubica la estación.

- Estación de la J.M.D.Chamberí, situada en Chamberí, colindante al distrito centro. Es el distrito con mayor densidad de habitantes, 29 364,3 hab/km^2. Destaca la escasez de parques, la variedad de puntos de interés para turistas, presencia de numerosas avenidas sin restricciones como el Paseo de la Castellana, que requieren de una gran cantidad de transporte público. Esto hace que se intuya como uno de los distritos en los que el efecto ICU sea más pronunciado.

- Estación de la J.M.D. Villaverde, situada en Villaverde. Junto a la estación de El Pardo se ubican más alejadas del centro que el resto. Se encuentra entre 2 parques, el Forestal Julio Alguacil Gómez y el de Plata y Castañar. Además, sin ser uno de los más extensos tiene una densidad de población no muy elevada y sumado a la poca influencia de carreteras, el efecto ICU se intuye menor.

- Estación de Plaza Elíptica, situada en Usera, entre la M-40 y M-30. Tiene presencia fluvial gracias al río Manzanares al este, el cual cruza un gran parque.

Además, con la presencia de otros parques como el Pradolongo, en este distrito predominan las zonas verdes.

- Estación de Cuatro Caminos, situada en la zona norte del centro de Madrid, en Tetuán. Tienen protagonismo los grandes edificios de oficinas en el Paseo de la Castellana, lo que lo hace que sea una zona tremendamente concurrida. El resto de su extensión, 5,37 km^2, son zonas residenciales, haciendo que con 153789 habitantes tenga una gran densidad. Sumado a su escasez de parques se espera de este distrito una mala calidad del aire.

- Estación de Escuelas Aguirre, situada en frente del Parque del Retiro. Esta estación cuenta con una superficie de 1,18 km^2 en la que predominan los árboles y el gran estanque interior. Rodeado por la calle O´Donnell con un gran tránsito de vehículos, la M-30, el centro de Madrid y la estación de Atocha. Esto hace que el posible efecto positivo del parque se vea reducido por la alta contaminación.

- Estación de El Pardo, al norte de la ciudad, es la estación más externa de la ciudad. Su extensión es la más elevada entre todos los distritos, 237,84 km^2. Y aunque es el segundo con mayor población, su densidad es la más baja de todas, 1003,86 hab/km^2. Además, salta a la vista la gran masa forestal que posee al norte, así como el gran Embalse del Pardo. Con esta información se presupone como uno de los distritos con mejor calidad de aire y menor efecto ICU de la ciudad.

- Estación del Juan Carlos I, en el distrito de Barajas, se encuentra en el segundo parque más grande de Madrid, 1,6 km^2, sólo después del parque forestal de Valdebebas. En su interior se encuentran grandes lagos, así como una gran masa arbórea lo que la temperatura dentro de él se encuentra regulada con respecto a los núcleos urbanos próximos.

Como ya hemos mencionado, hemos seleccionado estas estaciones por disponer de datos más completos. Dejando fuera estaciones no tan útiles para los objetivos. Aun así y viendo la Figura 2.2, se ve que las estaciones se distribuyen por todo el territorio.

2.1.2. Estaciones de Calidad del Aire

Tenemos las siguientes estaciones de calidad del aire de las que hemos extraído los datos de contaminación en la Figura 2.3

Ellas se dividen en:

- Interior M30 donde, siete son de tráfico (Escuelas Aguirre, Castellana, Plaza de Castilla, Ramón y Cajal, Cuatro Caminos, Plaza de España y Barrio del Pilar) más tres en segundo plano con respecto a las de tráfico (Plaza del Carmen, Méndez Álvaro y Retiro).

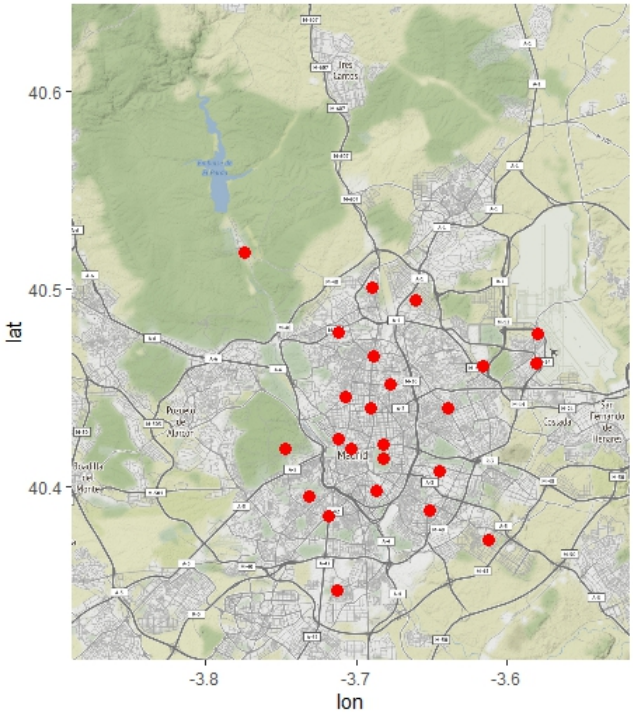


Figura 2.3: Imagen de elaboración propia

- Sureste de Madrid, una de tráfico (Moratalaz) más dos en segundo plano (Vallecas y Ensanche de Vallecas).
- Noreste de Madrid, cinco en segundo plano (Arturo Soria, Sanchinarro, Urbanización Embajada, Barajas Pueblo y Tres Olivos) más una suburbana (Juan Carlos I).
- Noroeste de Madrid, siendo dos suburbanas (El Pardo y Casa de Campo).
- Suroeste siendo una de tráfico (Plaza Elíptica) y dos en segundo plano (Farolillo y Villaverde).

2.2. Zonas verdes

En este punto hacemos hincapié sobre el efecto positivo de las zonas arbóreas en las temperaturas y más en las grandes urbes. Esto es lógico ya que los árboles no sólo tapan la luz solar reduciendo la temperatura ambiental, también suelen necesitar de un riego constante lo cual añade humedad al aire lo que mitiga el efecto del calor, sobre todo en temporada estival. Aun así, quisimos informarnos acerca de estos efectos. Y resulta que encontramos un artículo en el que se planteaba esta pregunta y el cómo afectaban las zonas verdes a la ICU de una ciudad. Arellano Ramos y Roca Cladera (2018) explica como la morfología de una ciudad afecta a su ICU, ya sean zonas verdes, fluviales y los mismos edificios con sus distintos materiales y distribución.

En concreto, Arellano y Roca obtuvieron como conclusiones que '*El factor clave son las diferencias en las estructuras verdes, teniendo un mayor alcance térmico debido a la mayor compacidad*', refiriéndose a los cambios de temperaturas entre los distritos analizados. Además de tener un '*mejor desempeño climático de Parc Central. Tanto de día como de noche*'. En esta Figura 2.4, obtenida de la Environmental Protection Agency (2008), citada en el artículo, podemos ver las diferencias de temperatura tanto en superficies como en el aire, en las distintas áreas dependiendo de si es de día o de noche.

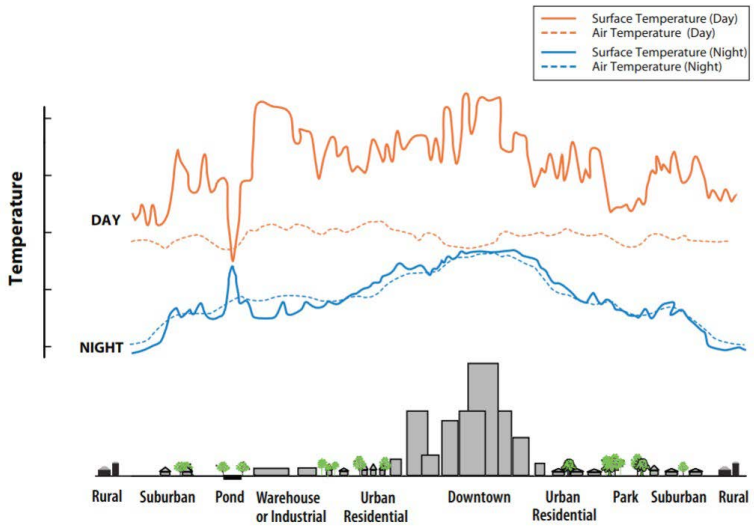


Figura 2.4: Imagen obtenida del artículo Arellano Ramos y Roca Cladera (2018)

Del mismo modo, este mismo estudio concluyó que se puede '*afirmar que la morfología urbana juega un papel clave; asimismo que la morfología urbana, el diseño del paisaje, la selección de la vegetación y de los materiales tienen especial relevancia en la resiliencia al cambio climático y en la resistencia a los eventos de ola de calor.*' Es por esto por lo que decidimos incluir en nuestro trabajo un análisis sobre todo el planeamiento urbano que tiene la ciudad de Madrid y más en concreto las zonas verdes y masas arbóreas.

2.3. Medidas meteorológicas

En este punto hacemos referencia a las distintas magnitudes que tendremos en cuenta en el análisis de los datos. Dichas magnitudes son recogidas por las estaciones meteorológicas de la Red de Calidad del Aire de Madrid[3].

2.3.1. Temperatura

Como ya hemos definido en puntos anteriores, el efecto *ICU* es principalmente una acumulación de calor. Nuestro análisis gira en torno a esta magnitud.

[3]Enlace Sistema Integral de Calidad del Aire

El clima de Madrid se caracteriza por las elevadas temperaturas en verano, y bajas en invierno. Es decir, suele haber una gran diferencia entre máximas y mínimas, en el año a analizar (2021) fueron 33ºC y 1ºC respectivamente, según Weather Spark. Esto supone una diferencia de 32º, que en comparación con ciudades como Barcelona 28º y 5º, Santiago de Compostela 24ºC y 5ºC, Sevilla 35ºC y 6ºC o Guadalajara 35ºC y 5ºC, es mayor aun estando a más altitud. Esto muestra que, el hecho de ser una gran urbe con las características que hemos mencionado en la introducción afecta en las temperaturas haciendo que la temperatura máxima sea extrema en comparación con su localización geográfica.

Como veremos en los resultados obtenidos, la diferenciación es mucho mayor cuando tratamos las diferentes regiones de Madrid. Siendo las zonas más céntricas mucho más afectadas por el calor y las temperaturas máximas en verano. Figura 2.5

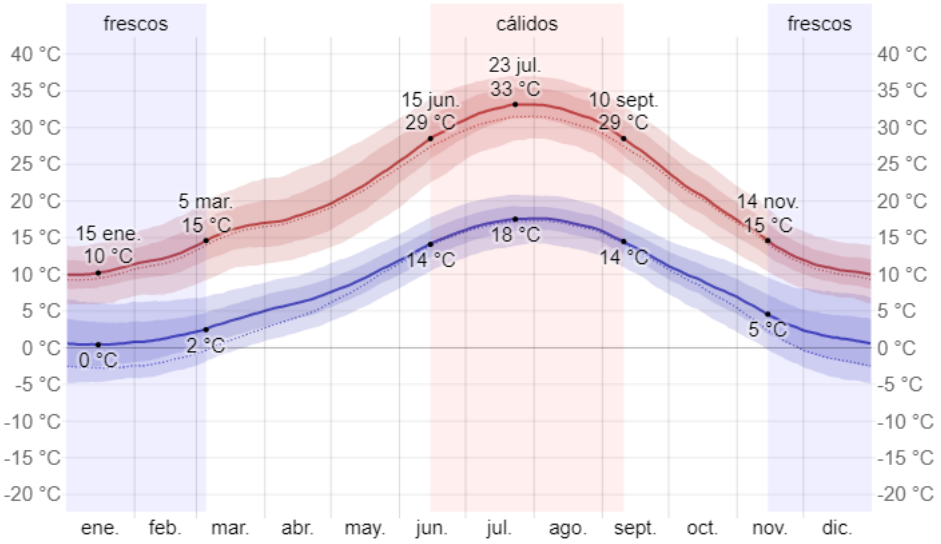


Figura 2.5: Imagen obtenida de Weather Spark, describe las temperaturas a lo largo del año 2021 en Madrid.

2.3.2. Humedad relativa

Esta magnitud mide, en porcentaje, la proporción entre la cantidad de humedad atmosférica presente y la cantidad que hubiese si el aire estuviera saturado. El aire caliente puede absorber más humedad que el aire frío. La humedad relativa indica cuánta humedad de lo que es físicamente posible está realmente contenida en el aire. Si la humedad es alta, la gente se siente incómoda y la encuentra opresiva. En general, una humedad relativa del 40-60 % se considera confortable. Con una humedad media del 78 %, es más desagradable en enero. En julio, en cambio, es más fácil de soportar. Figura 2.6 muestra la humedad relativa a lo largo del año 2021 en Madrid.

Puesto que el agua es uno de los mejores reguladores térmicos, a cuanta mayor humedad relativa implica unas temperaturas más suavizadas. Y aunque esto se puede experimentar al dar un paseo por un parque o estanque en la ciudad, hay artículos que lo confirman. Es decir, no sólo las precipitaciones influyen sobre esta

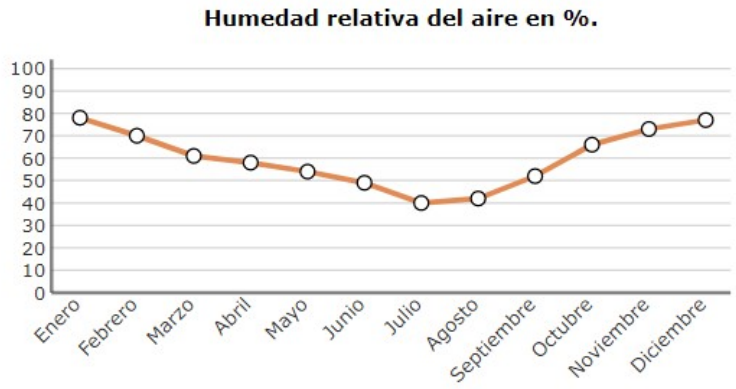


Figura 2.6: Imagen obtenida de DatosMundial.com, humedad relativa en Madrid.

magnitud, sino también las zonas verdes, pavimentos, etc. Tal y como se explica en Tejedor et al. (2016), saca como conclusión que: "*en los periodos cálidos la ciudad de Zaragoza agudiza de forma considerable el calor y la sequedad en relación a las zonas no urbanas.*". Es por esto por lo que incluir esta medida en la aplicación de los algoritmos nos ayudó a sacar conclusiones.

2.3.3. Precipitación

Esta magnitud influye directamente sobre la anterior. Si hay precipitación habrá una mayor cantidad de humedad atmosférica, lo que se traduce en un mayor porcentaje de humedad relativa.

Es cierto que Madrid no es una ciudad muy lluviosa, Figura 2.7, pero creímos oportuno tener en cuenta las precipitaciones para ver si a lo largo del año afectaba de igual manera a las estaciones céntricas que a las situadas en grandes parques o en el exterior de las circunvalaciones que rodean la ciudad.

Días de	ene.	feb.	mar.	abr.	may.	jun.	jul.	ago.	sept.	oct.	nov.	dic.
Lluvia	4,9d	4,2d	4,5d	6,2d	6,6d	3,9d	1,8d	2,0d	3,2d	6,0d	5,9d	5,2d

Figura 2.7: Imagen obtenida de Weather Spark, indica los días de lluvia en Madrid por mes en el año 2021.

2.3.4. Velocidad y dirección del viento

Dependiendo de la ubicación de la estación, el viento será un factor determinante. Ya que dependiendo de la topología de lo que le rodea ya sean edificios, árboles, carreteras, o simplemente en grandes parques abiertos, repercutirá sobre las magnitudes. Calles estrechas en las que haya más recovecos y la que las alturas de los edificios sean mayores, intensificarán el *efecto cañón*. Estas condiciones hacen que las calles atrapen el calor, reduciendo la posibilidad de que se disipe. Por lo que

la temperatura suele ser mayor que la media. Además de contribuir a niveles de contaminación más altos, Figura Lipp (2014).

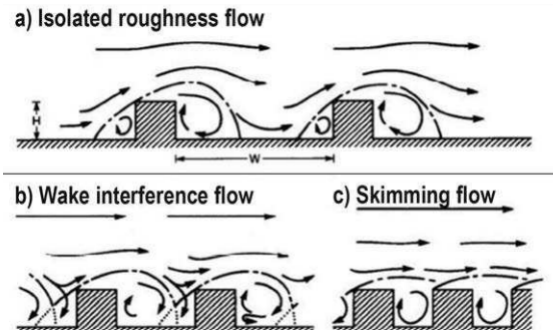


Figura 2.8: Imagen obtenida de Lipp (2014), muestra como el aire caliente queda atrapado entre edificios c).

La velocidad del viento afecta directamente a la temperatura. Esto se debe a que contribuye a la evaporación de la humedad en el ambiente aumentando la sensación de frío. Por lo que las temperaturas serán más bajas en lugares más amplios en los que haya una corriente continua y directa sobre la zona. O lo que es lo mismo, en zonas urbanas donde predominan edificios habrá una menor velocidad y por lo tanto una temperatura mayor.

La velocidad promedio en Madrid no se ve afectada entre estaciones, hay una cierta estabilidad a lo largo del año. Teniendo el fin del verano los vientos más calmados y el inicio de la primavera los más ventosos, como muestra la Figura 2.9.

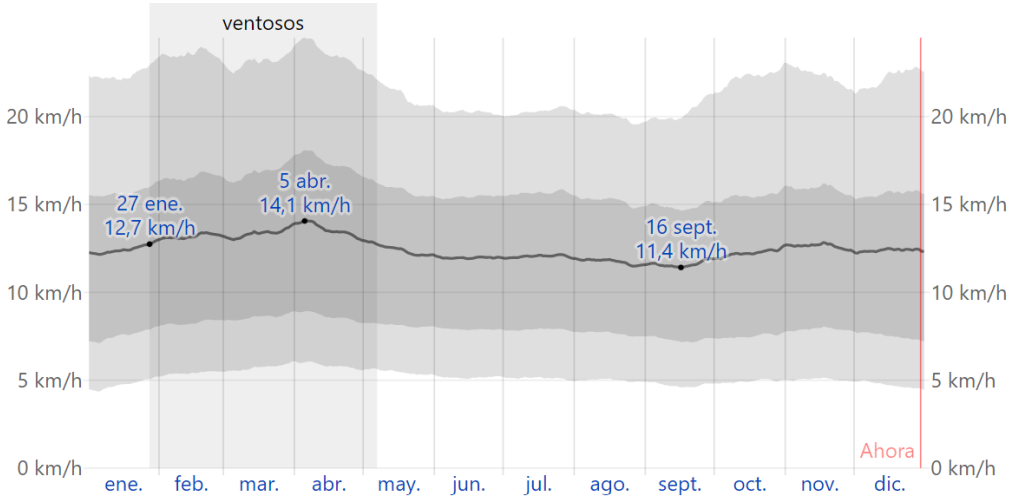


Figura 2.9: Imagen obtenida de Weather Spark, indica la velocidad del viento en Madrid por mes del año 2021.

Del mismo modo pasa con la dirección del viento. No hay una orientación que predomine ampliamente sobre el resto, al igual que con las velocidades. Aunque si hay una que ligeramente supera al resto es el visto de dirección Oeste, o cómo se

refiere en meteorología Céfiro, aunque muy igualado con el viento del Norte, como se ve en la Figura 2.10.

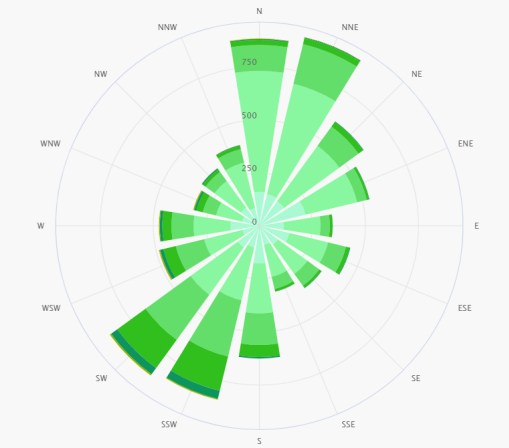


Figura 2.10: Imagen obtenida de Meteo Blue, indica la rosa de los vientos en Madrid el año 2021.

2.4. Proximidad a carreteras

Otro de los factores que influye en la subida de las temperaturas en las ciudades son las carreteras. Esto se debe principalmente a dos factores.

El primero y más obvio es la contaminación producida por vehículos. Es un hecho probado que el sistema de carreteras, autopistas y avenidas de una ciudad afecta en el calor de la misma, así como en la salud de los habitantes, Piracha y Chaudhary (2022).

El segundo y no tan claro son los materiales con los que están construidas todas estas carreteras. Refiriéndonos al asfalto y hormigón, son propensos a absorber y almacenar la radiación solar, Xu et al. (2021). Esto implica que por la tarde-noche, cuando las temperaturas deben bajar, se irradie el calor provocando que las temperaturas aumenten.

En la Figura 2.2 podemos ver la cercanía de las carreteras en las estaciones. Se puede observar la influencia de la M-40 y M-30 sobre las estaciones del Centro Municipal de Acústica, J.M.D. de Vallecas, Moratalaz, etc. Así como las grandes avenidas que cruzan la ciudad como el Paseo de la Castellana o la calle Alcalá afectan sobre las estaciones más céntricas.

2.5. Zonas azules

Otro de los factores influyentes en las temperaturas son las aguas. Ya sean estanques, lagos, fuentes ornamentales o simplemente el efecto del río.

La superficie de agua es otro elemento capaz de mitigar la isla de calor, Salinas 2020. Esto se confirma ya que, en la ciudad analizada en este estudio, Londres,

la diferencia térmica entre distritos que estaban más cerca de un corredor fluvial (gran masa de agua, en este caso el Támesis) y los más alejados o con menos masas, llegaba a ser de unos 3,5º-4,5ºC. Es decir, las zonas más próximas tenían unas temperaturas más regulares a lo largo del año. Figura 2.11

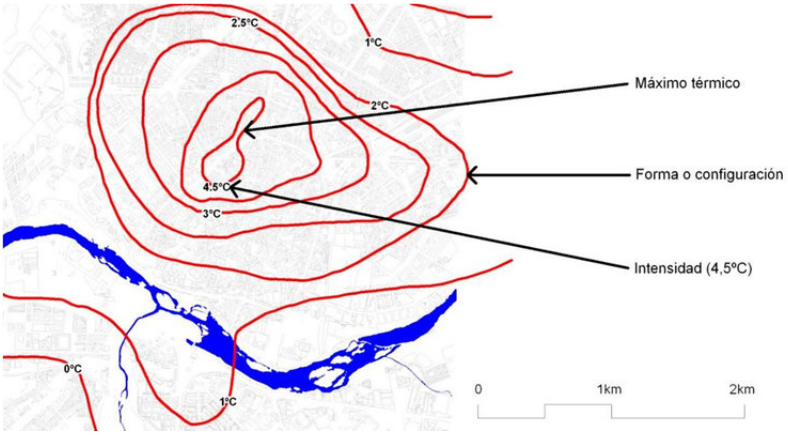


Figura 2.11: Intensidad térmica del agua. Fuente Salinas 2020.

En Madrid no abundan muchas de estas zonas. Es más, desde hace décadas el caudal del Río Manzanares a su paso por la ciudad se encuentra por debajo de la media en su curso alto y medio. Con una cifra de 12-15 m^3/s, es mucho menos que el caudal del Río Jarama que cruza la comunidad de Madrid por el oeste, con 31,7 m^3/s de media.

Además la ciudad dispone de pocas grandes masas de agua en el término municipal. Los estanques del Retiro, Parque Forestal de Valdebernardo, Parque Juan Carlos I, Parque Forestal Felipe VI, el Lago de Casa de Campo y sus arroyos. Todas estas masas están situadas fuera de núcleos urbanos, en parques, lo que aumenta todavía más las diferencias entre las temperaturas de estos núcleos y los parques.

Hay que nombrar también al embalse del Pardo, que con una capacidad de 45 hm^3 supone un gran alivio térmico para la región.

Todos estos factores empiezan a explicar las grandes diferencias observadas, como en el ejemplo de la Figura 1.1.

2.6. Medidas contaminantes

Gracias a estudios realizados previamente, Ballester (2005), se determina que para la salud humana y el impacto en el ecosistema los principales gases contaminantes serian: Dióxido de azufre (SO2), Monóxido de Carbono (CO), Óxidos de Nitrógeno (NO2, NO) y Ozono (O3); además, las partículas tienen en una amplia gama de tamaños y se clasifican en función de su diámetro aerodinámico en PM10 (partículas con un diámetro aerodinámico inferior a 10 micras) o PM2.5 (diámetro aerodinámico inferior a 2,5 micras). Por lo que definiremos estas y más adelante serán las que analicemos en el trabajo.

2.6.1. Dióxido de Azufre

Este contaminante tiene un periodo de vida de alrededor de 72 horas, y su principal fuente es la quema de combustibles fósiles ricos en azufre como es el petróleo, así como las erupciones volcánicas. Su inhalación repercute en afecciones del sistema respiratorio y en el funcionamiento de los pulmones. Asimismo, puede producir irritación ocular.

2.6.2. Monóxido de Carbono

Este contaminante tiene un periodo de vida de unos 30-90 días, su principal fuente es la quema de combustibles fósiles y biomasa. En la ciudad de Madrid viene derivado principalmente de los vehículos de combustión y maquinaria de climatización.

La exposición a este contaminante reduce la capacidad de transportar oxígeno de la sangre a los tejidos corporales.

2.6.3. Óxidos de Nitrógeno

Se pueden identificar varios tipos de Óxidos de Nitrógeno. El Monóxido de Nitrógeno (NO) es un tóxico que se oxida con rapidez convirtiéndose en Dióxido de Nitrógeno (NO2). El origen es tanto natural, debido a la descomposición bacteriana e incendios, como derivado de la actividad humana, debido a vehículos motorizados y quema de combustibles fósiles.

A su vez, el Dióxido de Nitrógeno resulta de la combustión efectuada a altas temperaturas, tanto de origen natural como antropogénico. Este tóxico, es irritante y precursor de la formación de contaminantes secundarios como el Ozono y las partículas menores de 2.5 micras.

Además, ambos (NOx) tienen un efecto corrosivo sobre la piel y el sistema respiratorio, pudiendo ser causante de un edema pulmonar cuando el sujeto se expone a concentraciones elevadas.

2.6.4. Ozono

Se genera por la reacción fotoquímica de los precursores NOx y CO, sustancias emitidas de forma directa que reaccionan con la luz solar en condiciones atmosféricas estables. Su impacto para la salud es notable, debido a que posee un elevado carácter oxidante que le proporciona la capacidad de destruir organismos completos.

2.6.5. Partículas Menores a 2,5 micras

La principal fuente de las partículas menores de 2,5 micras es la emisión producida por los vehículos diésel, derivados del tráfico rodado de las ciudades.

Su impacto sobre la salud está relacionado con enfermedades de tipo respiratorio, tales como la bronquitis y dolencias de tipo cardiovascular, además de provocar asma y alergias entre la población infantil. Gracias a su facilidad para ser respirado y su

reducido tamaño, puede viajar profundamente en los pulmones. Lo que lo hace que sea más perjudicial que las partículas menores de 10 micras, y tenga un mayor impacto en las personas.

2.6.6. Partículas Menores a 10 micras

Las fuentes de emisión de estas partículas pueden ser móviles o estacionarias, un 77,9 % del total procede del polvo suspendido existente en la atmósfera. La industria, la construcción y los comercios suponen el 7,6 %, mientras que el tráfico rodado supone un 6,5 %. La exposición prolongada puede provocar efectos nocivos en el sistema respiratorio. No obstante, como se ha mencionado anteriormente, es menos perjudicial que las partículas menores de 2.5 micras, ya que al tener un mayor tamaño estas no logran atravesar los alvéolos pulmonares, quedando retenidas en la mucosa que recubre las vías respiratorias, siendo expulsadas de manera relativamente eficaz mediante la tos.

Capítulo 3

Origen y obtención de los datos

En este punto del proyecto pretendemos explicar con detalle la procedencia de los datos, los cuales deben ser lo más fiables y veraces posibles ya que las conclusiones finales del trabajo dependen principalmente de dichos datos. Del mismo modo trataremos de indicar paso a paso como han sido obtenidos. En nuestro caso estos datos serán meteorológicos, contaminación y zonas verdes. Todos ellos serán explicados con detenimiento en su punto correspondiente. El objetivo de la obtención de estos datos es principalmente el tratamiento de ellos, limpieza, preparación, establecer un formato, para posteriormente poder procesar y sacar conclusiones.

3.1. Origen y formato de los datos

En cada apartado indicaremos el origen para poder acceder a una copia original de los datos tomados. Principalmente el origen de los datos es el **Portal de datos abiertos del Ayuntamiento de Madrid**[1] que es un catálogo de conjuntos de datos recogidos y publicados por el Ayuntamiento de la capital. Cada conjunto de datos del portal se puede descargar en distintos formatos, siendo los principales CSV, TXT y XML. En nuestro caso utilizaremos CSV. Y se encuentran todos recogidos en el repositorio GitHub. Además, el portal proporciona la fecha a partir de la cual se empezaron a recolectar los datos, así como la fecha de incorporación al catálogo, palabras claves para poder filtrar entre los conjuntos de datos, el responsable del conjunto de datos, editor, lugar y licencia de todos los datos.

3.1.1. Datos meteorológicos

La ciudad de Madrid cuenta con un **Sistema Integral de la Calidad del Aire**[2] que consta de una red meteorológica, con estaciones de este tipo por cada municipio. Desde el portal del SICA podemos ver cómo están repartidos geográficamente las distintas estaciones. Estas, son capaces de medir diferentes magnitudes, explicadas a través de un intérprete que está a disposición del público, que resumidamente son:

[1] Portal de datos abiertos
[2] Calidad del aire de Madrid. SICA

Código	Parámetro	Unidad de medida
80	RADIACIÓN ULTRAVIOLETA	Mw/m2
81	VELOCIDAD VIENTO	m/s
82	DIR. DE VIENTO	º
83	TEMPERATURA	ºC
86	HUMEDAD RELATIVA	%
87	PRESIÓN BAROMÉTRICA	mb
88	RADIACIÓN SOLAR	W/m2
89	PRECIPITACIÓN	l/m2

Tabla 3.1: Datos meteorológicos

De igual modo para poder identificar cada una de las estaciones nos referiremos a la siguiente relación código-estación, refiriéndose en algunos casos a la estación como los últimos tres dígitos del código. Tabla 3.2

El formato del conjunto de datos diario se puede observar en la Tabla 3.3

Se especifican un periodo de análisis y una técnica de muestreo, que son las que implementan las estaciones que miden las variables meteorológicas. Junto al valor recogido por cada día, hay una columna que hace referencia a su validación. Siendo V válido y N no válido. En este formato, dependiendo del mes habrá columnas vacías (p.e. Septiembre tendrá D31 a 0 y V31 a N).

3.1.2. Localización y sensores de las estaciones

Del mismo portal de datos abiertos obtuvimos un archivo .xls en el que se asocia cada estación a su localización.

En este archivo se encuentran las coordenadas en numerosos formatos, lo que nos ayudó a conseguir las representaciones de una forma más sencilla. Además, este archivo recoge los sensores de los que dispone cada estación ya que no todas tienen los mismos.

El formato tiene los siguientes valores:

CÓDIGO, CÓDIGO_CORTO, ESTACIÓN, DIRECCIÓN, LONGITUD_ETRS89, LATITUD_ETRS89, ALTITUD, (81), DV (82), T (83), (86), PB (87), (88), (89).

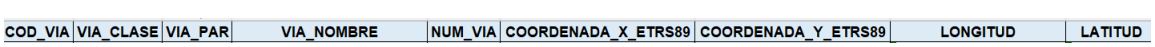


Figura 3.1: Formato estaciones de control

En cuanto a las estaciones y sus sensores, la Red de meteorología del Ayuntamiento[3] los recoge en este link[4].

[3]https://www.madrid.es/portales/munimadrid/es/Medio-Ambiente/Red-de-meteorologia

[4]Link a las estaciones y sus sensores

Código	Estación
28079102	J.M.D. Moratalaz
28079103	J.M.D. Villaverde
28079104	E.D.A.R. La China
28079106	Centro Mpal. De Acústica
28079107	J.M.D. Hortaleza
28079108	Peñagrande
28079109	J.M.D.Chamberí
28079110	J.M.D.Centro
28079111	J.M.D.Chamartín
28079112	J.M.D.Vallecas 1
28079113	J.M.D.Vallecas 2
28079114	Matadero 01
28079115	Matadero 02
28079004	Plaza España
28079008	Escuelas Aguirre
28079016	Arturo Soria
28079018	Farolillo
28079024	Casa de Campo
28079035	Plaza del Carmen
28079036	Moratalaz
28079038	Cuatro Caminos
28079039	Barrio del Pilar
28079054	Ensanche de Vallecas
28079056	Plaza Elíptica
28079058	El Pardo
28079059	Juan Carlos I

Tabla 3.2: Tabla estaciones meteorología

3.1.3. Datos de contaminación

Al igual que en el punto anterior vamos a utilizar el Sistema Integral de la Calidad del Aire para recoger los datos de contaminación. Las diferentes magnitudes que es capaz de medir serían las que muestra la Tabla 3.4:

Las medidas de contaminantes que nos proporcionan las estaciones son las siguientes:

De igual modo para poder identificar cada una de las estaciones nos referiremos a la siguiente relación código-estación, refiriéndose en algunos casos a la estación como los últimos dos dígitos del código, como se ve en la Tabla 3.5.

En cuanto al formato de estos datos es prácticamente el mismo que el de meteorología. Difieren en que no indican ni provincia, municipio, ni técnica. Pero en cuanto al valor y su campo de validación sigue el mismo funcionamiento.

Por lo tanto el formato de los datos es el que muestra la Tabla 3.6.

Prov.	Mun.	Estación	Magnitud	Técnica	Periodo	Año	Mes	D01	V01
28	79	54	81	98	02	2022	7	1.18	V

Tabla 3.3: Tabla datos diarios

Magnitud	Parámetro	Unidad de medida
01	Dióxido de Azufre	SO2
06	Monóxido de Carbono	CO
07	Monóxido de Nitrógeno	NO
08	Dióxido de Nitrógeno	NO2
09	Partículas <2.5 um	PM2.5
10	Partículas <10 um	PM10
12	Óxidos de Nitrógeno	NOx
14	Ozono	O3
20	Tolueno	TOL
30	Benceno	BEN
35	Etilbenceno	EBE
37	Metaxileno	MXY
38	Paraxileno	PXY
39	Ortoxileno	OXY
42	Hidrocarburos totales	TCH
43	Metano	CH4
44	Hidrocarburos no metánicos	NMHC

Tabla 3.4: Tabla Contaminación

3.1.4. Datos Zonas verdes

Para ello recurrimos de nuevo al Portal de datos abiertos del Ayuntamiento de Madrid, específicamente al conjunto **Masas arbóreas en zonas verdes/distrito**[5], donde se realiza una asociación entre distrito y la masa arbórea que el propio Ayuntamiento ha contabilizado. Indica el número de masas arbóreas que se pueden corresponder desde jardines hasta parques forestales.

La masa arbórea se nos proporciona en dos magnitudes, m^2 y ha.

El formato se ve en la Tabla 3.7:

[5]Arboleda en Madrid

Código	Estación
28079004	PZA. DE ESPAÑA
28079008	ESCUELAS AGUIRRE
28079011	AV. RAMÓN Y CAJAL
28079016	ARTURO SORIA
28079017	VILLAVERDE ALTO
28079018	C/ FAROLILLO
28079024	CASA DE CAMPO
28079027	BARAJAS
28079035	PLAZA DEL CARMEN
28079036	MORATALAZ
28079038	CUATRO CAMINOS
28079039	BARRIO DEL PILAR
28079040	PUENTE DE VALLECAS
28079047	MENDEZ ALVARO
28079048	CASTELLANA
28079049	RETIRO
28079050	PLAZA CASTILLA-CANAL
28079054	ENSANCHE DE VALLECAS
28079055	URBANIZACIÓN EMBAJADA
28079056	PLAZA ELIPTICA
28079057	SANCHINARRO
28079058	EL PARDO
28079059	JUAN CARLOS I
28079060	TRES OLIVOS
28079061	CENTRO INTEGRADO ARGANZUELA

Tabla 3.5: Tabla de las estaciones contaminación

Estación	Magnitud	Año	Mes	D01	V01
54	8	2022	7	10	V

Tabla 3.6: Formato de la tabla de contaminación

Distrito	Nº de masas arbóreas	Superficie masa arbórea(m)	Superficie masa arbórea(ha)

Tabla 3.7: Formato de la tabla de masa arborea

Capítulo 4

Preprocesado de datos

En este capítulo explicamos y detallamos los pasos que hemos dado para obtener un conjunto de datos útil. Este nos servirá en puntos posteriores para un análisis y obtención de conclusiones sobre el efecto ICU en Madrid.

Todos los archivos que mencionaremos de aquí en adelante se encuentran en el repositorio de **GitHub**[1]. Tanto scripts en R, como los CSV´s y Excel´s usados.

4.1. Conjuntos de datos

Primero descargamos los ficheros sobre las estaciones de calidad del aire y las estaciones meteorológicas proporcionados por el Ayuntamiento de Madrid. Este paso se ha detallado en el capítulo anterior.

En el proyecto decidimos elegir el siguiente formato del directorio de carpetas y subcarpetas, que incluyen los datos limpiados, los de elaboración propia y las aproximaciones que obtenemos en el capítulo 6:

- Meteo
  - Limpiar: meteo21.csv: es el conjunto ya citado que recoge todos los datos meteorológicos y que disponemos a limpiar y formatear.
  - estacion_distrito.csv: este archivo es de elaboración propia, y asignamos cada estación de meteorología a su distrito y añadimos su nombre.
  - Estaciones_control_datos_meteorologicos.xls: ya citado en el punto anterior.
  - Diario: aquí guardamos la evolución de los .csv que tratamos, desde el primero que descargamos hasta el último a analizar.
- Contaminación, que sigue el mismo formato que Meteo.
  - Limpiar: contaminacion21.csv: es el conjunto ya citado que recoge todos los datos de calidad del aire y que disponemos a limpiar y formatear.

[1] https://github.com/Gmene00/TFG-Madrid-Isla-de-calor

  - estacion_distrito.csv: este archivo es de elaboración propia, y asignamos cada estación de calidad del aire a su distrito y añadimos su nombre.
  - informacion_estaciones_red_calidad_aire.xls: ya citado en el punto anterior.
  - Diario: aquí guardamos la evolución de los .csv que tratamos, desde el primero que descargamos hasta el último a analizar.

- Zonas Verdes

  En la raíz de esta carpeta se encuentran los archivos después de ser tratados.

  - Limpiar: es el conjunto de .csv que recoge todos los datos de masa arbórea y zonas verdes que disponemos a limpiar y formatear.

- Mapa

  - Clustering: Guardamos las aproximaciones realizadas para su representación gráfica en un mapa.

- Datos Finales

  Aquí guardamos los .csv de los datos ya limpiados y depurados para los análisis.

- En la raíz tenemos todos los archivos .R que hemos utilizado en el proyecto.

4.2. Limpieza de datos

Este apartado es relevante en el trabajo y ha supuesto la mayor carga de trabajo tanto por esfuerzo como por tiempo, donde hemos conseguido simplificar todos los datos a un conjunto más fácil de utilizar y de interpretar.

Antes de la limpieza recabamos los datos disponibles por estación y magnitud. Tras analizar los datos del ayuntamiento, vemos que hay estaciones que no recogen todas las magnitudes o no la recogen de forma completa ya que para ciertos meses no existen los valores debido a fallo de la estación o porque su validación fue N, haciendo referencia a *no*.

En las Figuras 4.1 y 4.4, podemos observar los datos que nos proporciona el Ayuntamiento y cuyo contenido se explica a continuación.

- Sí: disponemos de todos los datos, a excepción de algún día puntual que no fuese válido.

- No: el Ayuntamiento no lo proporciona y por tanto **no** aparecen en el conjunto de datos.

- A medias: disponemos de los datos parcialmente, es decir, hubo un fallo prolongado en el tiempo, mes/meses por lo que hay una mayor cantidad de valores no válidos. Estos sí aparecen, pero con el bit de validación a N.

	RADIACIÓN ULTRAVIOLETA	VELOCIDAD VIENTO	DIR. DE VIENTO	TEMPERATURA	HUMEDAD RELATIVA	PRESIÓN BAROMÉTRICA	RADIACIÓN SOLAR	PRECIPITACIÓN
	80	81	82	83	86	87	88	89
J.M.D. Moratalaz	NO	SI	SI	SI	SI	SI	SI	SI
J.M.D. Villaverde	NO	SI	SI	SI	SI	SI	SI	SI
E.D.A.R. La China	NO	SI	SI	NO	NO	NO	NO	NO
Centro Mpal. De Acústica	NO	SI	SI	SI	SI	SI	SI	SI
J.M.D. Hortaleza	NO	SI	SI	SI	SI	SI	SI	SI
Peñagrande	NO	A MEDIAS	A MEDIAS	A MEDIAS	A MEDIAS	A MEDIAS	A MEDIAS	A MEDIAS
J.M.D.Chamberí	NO	NO	NO	SI	SI	NO	NO	NO
J.M.D.Centro	NO	NO	NO	SI	SI	NO	NO	NO
J.M.D.Chamartín	NO	NO	NO	A MEDIAS	A MEDIAS	NO	NO	NO
J.M.D.Vallecas 1	NO	NO	NO	SI	SI	NO	NO	NO
J.M.D.Vallecas 2	NO	NO	NO	SI	SI	NO	NO	NO
Matadero 01	NO	NO	NO	A MEDIAS	A MEDIAS	NO	NO	NO
Matadero 02	NO	NO	NO	A MEDIAS	A MEDIAS	NO	NO	NO
Plaza España	NO	NO	NO	A MEDIAS	NO	NO	NO	NO
Escuelas Aguirre	NO	NO	NO	SI	SI	NO	NO	NO
Arturo Soria	NO	NO	NO	NO	SI	NO	NO	NO
Farolillo	NO	NO	NO	SI	NO	NO	NO	NO
Casa de Campo	NO	SI	SI	SI	SI	A MEDIAS	SI	SI
Plaza del Carmen	NO	NO	NO	SI	SI	NO	NO	NO
Moratalaz	NO	NO	NO	SI	SI	NO	NO	NO
Cuatro Caminos	NO	NO	NO	SI	SI	NO	NO	NO
Barrio del Pilar	NO	NO	NO	NO	SI	NO	NO	SI
Ensanche de Vallecas	NO	SI	SI	SI	SI	NO	SI	SI
Plaza Elíptica	NO	SI	SI	SI	SI	SI	NO	SI
El Pardo	NO	NO	NO	SI	SI	NO	NO	NO
Juan Carlos I	NO	SI	SI	SI	SI	SI	SI	SI

Figura 4.1: Datos disponibles sobre meteorología.

Hay que señalar que hay varias estaciones que no recogen una gran parte de los datos, debido a que no tienen implementados los sensores. Vienen explicado en los siguientes .csv que podemos encontrar en las carpetas correspondiente:

- información_estaciones_red_calidad_aire.xls en la carpeta **contaminación**.
- Estaciones_control_datos_meteorológicos.xls en la carpeta **meteo**.

El Portal de la Red de Meteorología del Ayuntamiento recoge en su web estos dos archivos mencionados.

En base a estos datos vamos a realizar la limpieza. Tanto en el conjunto de meteorología como en el de contaminación hemos usado el mismo algoritmo de tratamiento de datos. Esto se debe a que su formato es prácticamente idéntico, por lo que con modificar el script de limpieza de los datos meteorológicos obtenemos la limpieza de los de contaminación.

Estos scripts se componen de estos pasos:

- Empezamos con una iteración sobre las filas de los conjuntos. Una fila representa los valores de **una** magnitud recogidos por **una** estación en **un** mes determinado. Cada una de ellas contiene 31 valores y 31 validaciones asociadas a cada uno de los valores. Por lo tanto, el siguiente paso es comprobar la validez de cada columna.

- Empezamos una segunda iteración. Por cada una de las filas nombradas recorremos sus datos y nos quedamos con los valores de los días. En el caso de que no sean válidos lo rellenamos con un NA. Al final de estas dos iteraciones conseguimos una validación de todos los datos y reducimos el tamaño del conjunto de datos. En este punto grabamos este data frame intermedio en un .csv.

- De las estaciones que faltan datos, A MEDIAS, los rellenamos a NA y volvemos a grabar los datos.

- En este paso es donde le damos un formato más visible y fácil de entender. Generamos por cada estación 365 filas correspondientes a los días del año, evitando así que aparezcan meses como febrero con 31 días, que era lo que se nos proporcionaba desde el Ayuntamiento. Cada una de estas filas recoge todas las magnitudes del conjunto de datos. Dando como resultado un nuevo .csv de 365*(número de estaciones del conjunto) filas y (número de magnitudes del conjunto)+5 columnas (que contienen la fecha e identificador de la estación). En la Figura 4.2 se ve el nuevo formato, más coherente e intuitivo.

Fecha	Año	Mes	Día	Estación	81	...	89
2021-01-01	2021	01	01	102	2,21	...	0,1
...	...	...	...	...	...	...	...
2021-12-31	2021	01	01	102	0,92	...	0

Figura 4.2: Formato tras limpieza.

- Por último, fuimos recogiendo en una tabla los valores nulos por cada uno de los conjuntos, para un posterior análisis de su distribución.

Tras esta descripción de cómo se ha hecho la limpieza vamos a indicar observaciones específicas por cada conjunto.

4.2.1. Limpieza de datos meteorológicos

Como ya hemos visto en la Figura 4.1, vemos diferencia de recogida de datos entre estaciones. Percibimos que tienen más sensores las ubicadas en parques, zonas verdes, etc. Por ejemplo, el efecto cañón, estudiado en el Estado de la Cuestión, que se produce en las zonas urbanas e influye directamente en el efecto ICU, se relaciona con la velocidad del viento entre otras variables. Este efecto se recoge en pocas estaciones urbanas.

Esto ocurre en más de la mitad de las estaciones, y concretamente, de las veintiséis, solo en ocho disponen de más de la mitad de los sensores. En el resto solo disponemos de dos o tres sensores.

Para poder verlo de una forma más visual nos apoyamos en la Figura 4.3 que muestra las estaciones. De rojo aquellas de las que sí se disponen más de la mitad de los sensores y de negro las que tienen menos.

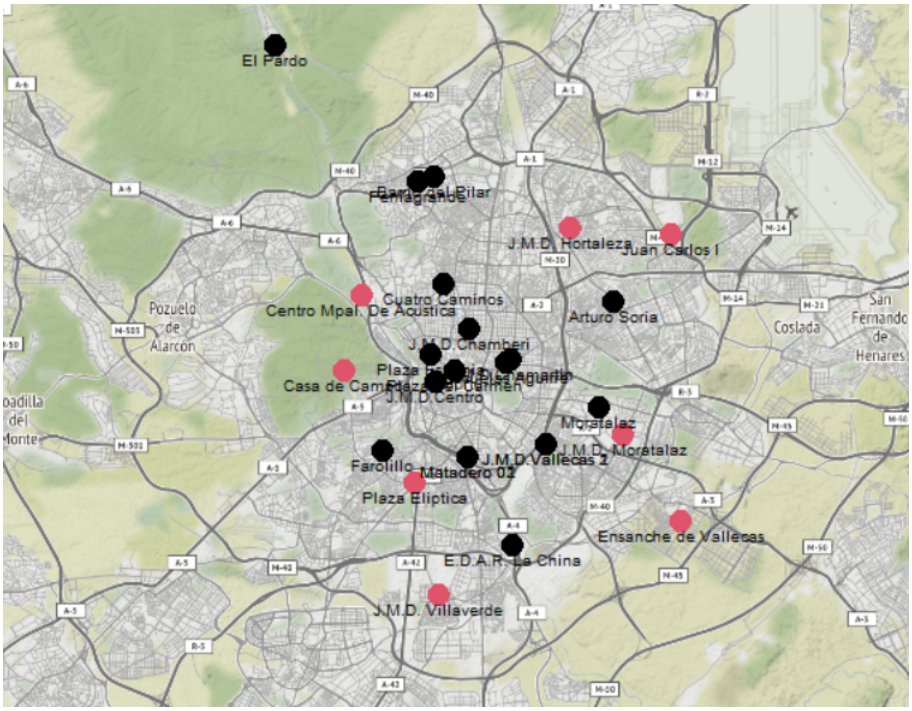


Figura 4.3: Imagen de elaboración propia, muestra las estaciones de las que se tienen más datos disponibles.

Otra observación destacable es que no tenemos ningún dato de la magnitud "80", radiación ultravioleta. Por el mismo motivo de no tener instalado el sensor en las estaciones. Y aunque el archivo hace referencia a esta magnitud, en el Portal del Ayuntamiento no aparece.

4.2.2. Limpieza de datos de contaminación

Lo primero es rechazar todas las magnitudes que no nos interesan para el estudio. Nos quedamos con los que afectan a la ciudad de Madrid coincidiendo con los principales gases contaminantes. En el análisis quitaremos alguna magnitud de acuerdo con lo concluido en la Memoria anual 2021 Calidad del Aire. Donde *"Se ha reducido el número de analizadores de dióxido de azufre y monóxido de carbono debido a las muy bajas concentraciones de estos contaminantes registradas en toda la ciudad. Los niveles obtenidos son inferiores al umbral inferior de evaluación que establece la legislación y que se define como "nivel por debajo del cual es posible limitarse al empleo de técnicas de modelización para evaluar la calidad del aire"*[2].

Del mismo modo que ha pasado con los datos de meteorología, hay muchas estaciones con pocos sensores como muestra la Figura 4.4.

[2]Memoria anual 2021 Calidad del Aire

	Dioxido de Azufre SO2	Dioxido de Nitrogeno NO2	Particulas < 2.5 um PM2_5	Particulas < 10 um PM10	Ozono O3
	1	8	9	10	14
PZA. DE ESPAÑA	A MEDIAS	A MEDIAS	NO	NO	NO
ESCUELAS AGUIRRE	SI	SI	SI	SI	SI
AV. RAMÓN Y CAJAL	NO	SI	NO	NO	NO
ARTURO SORIA	NO	SI	NO	NO	SI
VILLAVERDE ALTO	A MEDIAS	SI	NO	NO	SI
C/ FAROLILLO	NO	SI	NO	SI	SI
CASA DE CAMPO	A MEDIAS	SI	SI	SI	SI
BARAJAS	NO	SI	NO	NO	SI
PLAZA DEL CARMEN	SI	SI	NO	NO	SI
MORATALAZ	SI	SI	NO	SI	NO
CUATRO CAMINO	A MEDIAS	SI	SI	SI	NO
BARRIO DEL PILAR	NO	SI	NO	NO	SI
PUENTE DE VALLECAS	A MEDIAS	SI	NO	SI	NO
MENDEZ ALVARO	NO	SI	SI	SI	NO
CASTELLANA	NO	SI	SI	SI	NO
RETIRO	NO	SI	NO	NO	SI
PLAZA CASTILLA-CANAL	NO	SI	SI	SI	NO
ENSANCHE DE VALLECAS	NO	SI	NO	NO	SI
URBANIZACION EMBAJADA	NO	SI	NO	SI	NO
PLAZA ELIPTICA	NO	SI	SI	SI	NO
SANCHINARRO	SI	SI	SI	SI	NO
EL PARDO	NO	SI	NO	NO	SI
JUAN CARLOS I	NO	SI	NO	NO	SI
TRES OLIVOS	NO	SI	NO	SI	SI

Figura 4.4: Datos disponibles sobre Contaminación.

4.2.3. Limpieza de datos Estado Zonas Verdes de Distritos y Calles

Puesto que el formato de estos datos es una tupla que contiene el distrito, masas arbóreas, superficie y número de ejemplares, lo único que hay que realizar es eliminar la leyenda que incluyen desde el Ayuntamiento para informar sobre los campos.

4.3. Resumen

Al final de todo este proceso conseguimos 4 archivos, que se recogen en la carpeta **Datos Finales** del GitHub:

- meteoDiario21_byDay.csv: con los datos de meteorología con el nuevo formato legible y con datos validados.
- contaminaDiario21_byDay.csv: lo mismo que el punto anterior, pero tratándose de los datos de contaminación.
- masaArborea21.csv: datos de masas arbóreas limpios.
- datos21_byDay.csv: mediante un join juntamos ambos ficheros.

También se incluyen los ficheros que contienen las ubicaciones y medidas de las estaciones de meteorología y calidad del aire. Además de añadir las aproximaciones que desarrollamos en el capítulo 6.

4.3.1. Diagrama de flujo

Para poder ver todo este proceso de una forma más esquematizada y clara recurrimos a la Figura 4.5

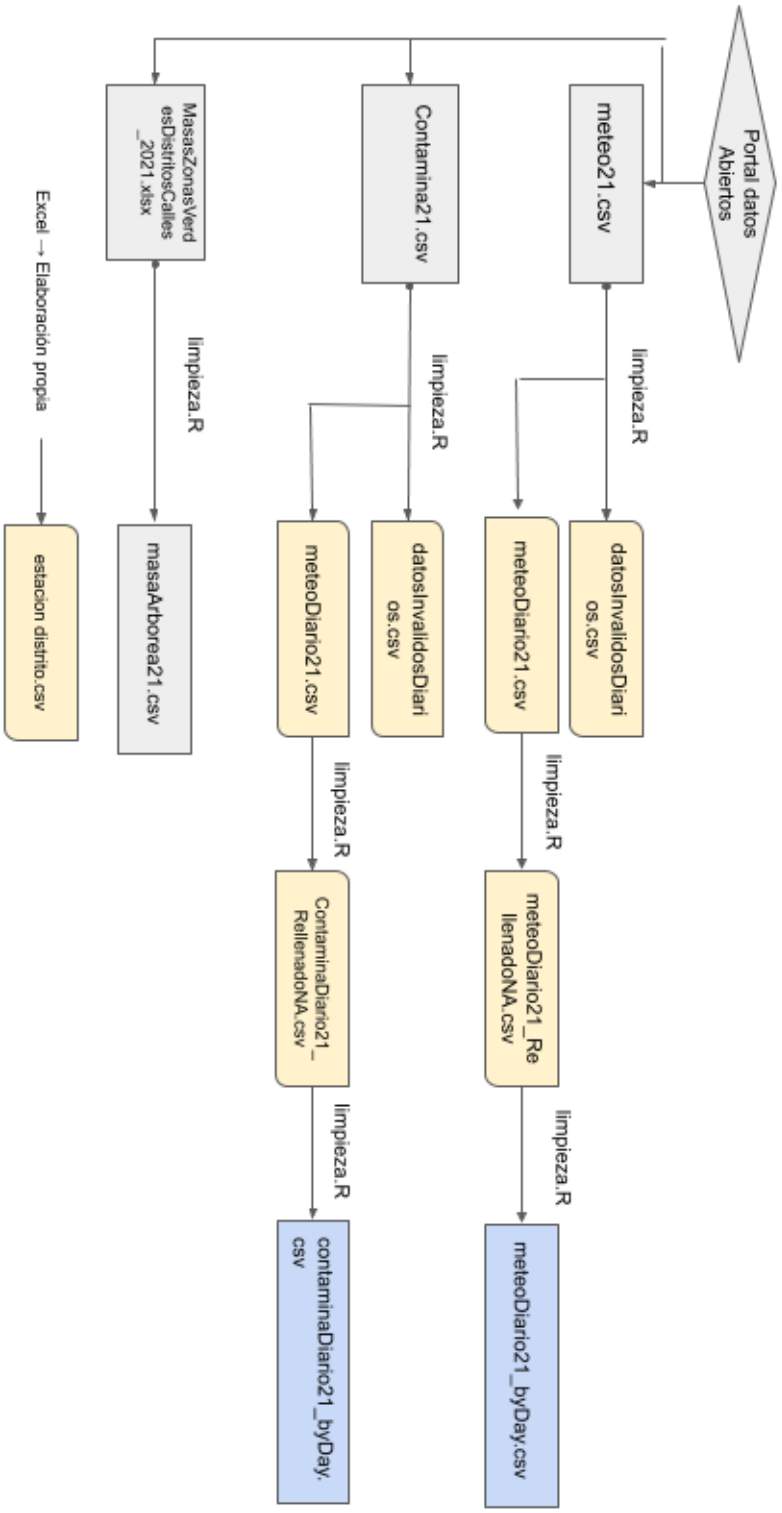


Figura 4.5: Diagrama de flujo de la limpieza

Capítulo 5

Estudio de los datos

En este capítulo, el objetivo es eliminar valores nulos y outliers, en la medida de lo posible para obtener un conjunto de datos consistente y listo para usar, logrando así los objetivos propuestos.

Vamos a analizar la calidad de los datos, ver cuántos y como se distribuyen los valores nulos, así como los atípicos y establecer en base a esto los datos finales a utilizar. Al sacar sus diagramas de densidad podemos ver la distribución que tienen, lo cual nos puede ser de utilidad esto lo vamos a analizar individualmente. De los datos de calidad del aire tenemos las siguientes distribuciones.

5.1. Distribución de los datos

Primero vemos en la distribución de los valores de dióxido de azufre que en general están comprendidos entre uno y cinco, teniendo un repunte en ocho. Figura 5.1a
En la distribución de la cantidad de Dióxido de nitrógeno tenemos los valores mayores comprendidos entre cero y setenta. Figura 5.1b

En la distribución de PM 2.5 tenemos los valores entre cero y quince. Figura 5.1c

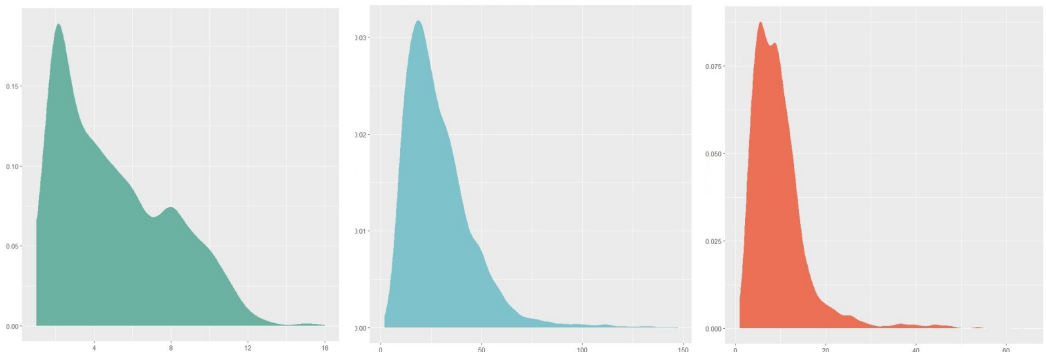
En la distribución de PM 10 tenemos los valores entre cero y veinticinco. Figura 5.2a

La distribución del ozono es claramente una campana de Gauss. Figura 5.2b

De los datos meteorológicos tenemos las siguientes distribuciones que a continuación pasamos a analizar.

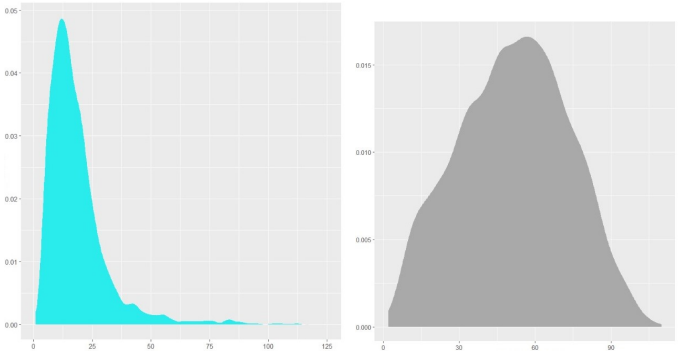
La mayoría de los datos se encuentran comprendidos en velocidades menos a 4 m/s que se corresponden correctamente con la ciudad de Madrid, Figura 5.3a, en este caso hay un mayor viento en dirección norte, la mayoría de empresas y sector industrial se encuentran al sur de la ciudad de Madrid, lo que puede provocar que traiga aire de mala calidad a la ciudad, Figura 5.3b donde 0^{o} seria dirección Norte y 180^{o} seria dirección Sur.

Dentro de los valores extremos que vamos a estudiar, podemos ver que en un



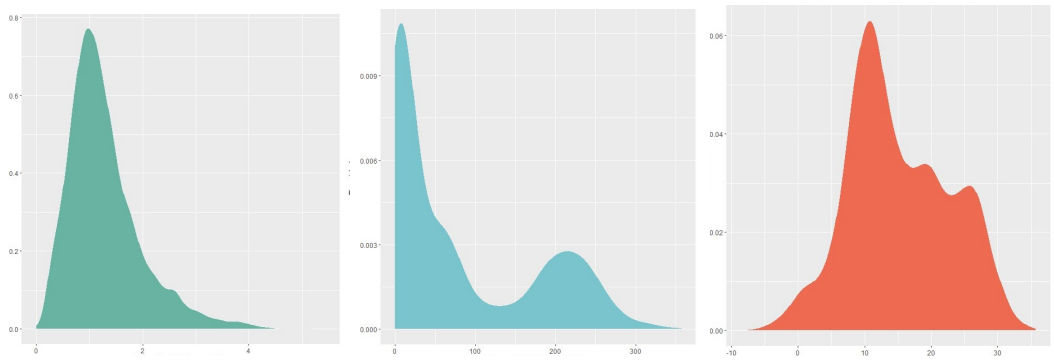
(a) Dióxido Azufre (b) Dióxido Nitrógeno (c) PM25

Figura 5.1: Distribución contaminantes 1



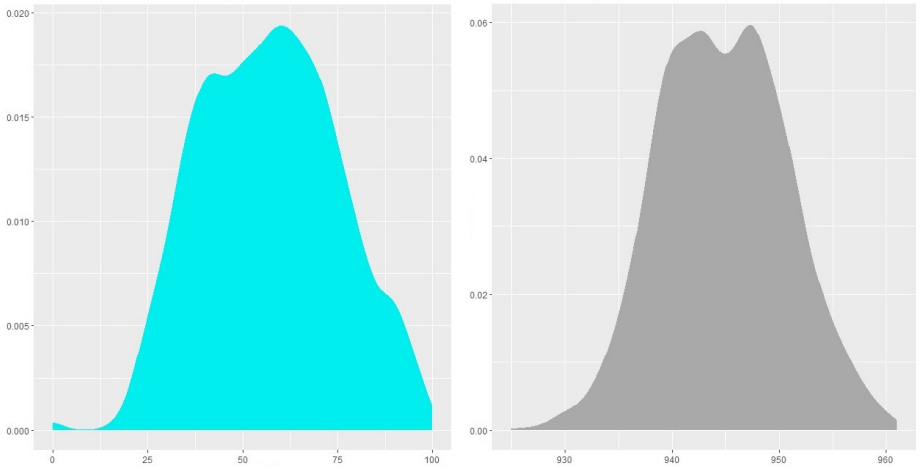
(a) Distribución PM10 (b) Distribución Ozono

Figura 5.2: Distribución contaminantes 2



(a) Velocidad viento (b) Dirección viento (c) Temperatura

Figura 5.3: Distribución meteorológica 1



(a) Humedad (b) Presión

Figura 5.4: Distribución meteorológica 2

60 % de los datos se encuentran en torno a los 10º. Figura 5.3c

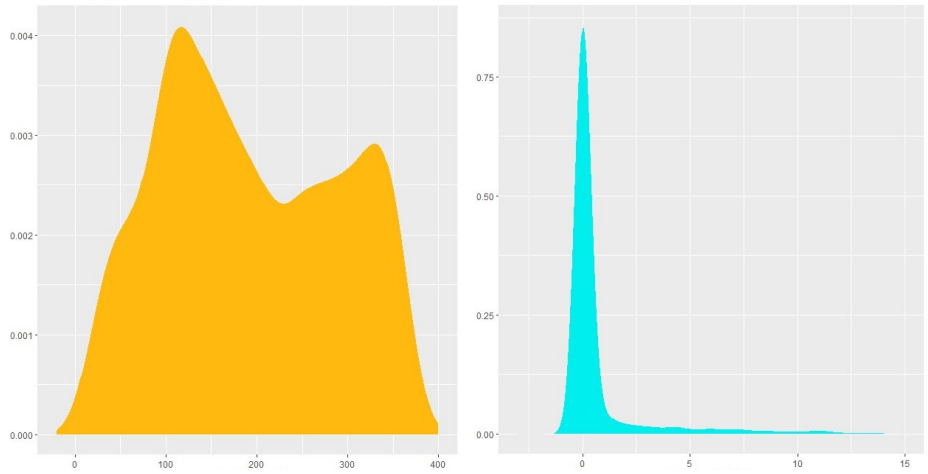
Podemos ver que la humedad promedio relativa de Madrid en 2021 fue en torno al 55 %, Figura 5.4a. Recordar que la humedad relativa comprendida entre 40-60 % se considera confortable. La presión del año 2021 se mueve mayormente entre los 940 y 950 mbar. Figura 5.4b

Entre los 100 y 200 W/m2 y los 300 y 350 W/m2 es donde se encuentran la mayor cantidad de datos sobre radiación. Figura 5.5a

Aquí podemos comprobar que las precipitaciones cuando ocurren no son muy abundantes, obteniendo pocos datos de más de 2 L/m2 en el año 2021. Figura 5.5b

5.2. Valores nulos en datos meteorológicos

Como ya mencionamos en el punto de la obtención de los datos, los archivos del portal de datos abiertos del ayuntamiento proporcionan por cada columna de valores, correspondientes a cada uno de los días del mes, un campo de validación.



(a) Radiación (b) Precipitación

Figura 5.5: Distribución meteorológica 3

Para poder poner en contexto de una forma más rápida los datos decidimos crear una tabla con las magnitudes de las que disponían cada una de las estaciones, Figura 4.1.

Aparte hemos creado la Tabla 5.6 donde vemos los meses que están a medias y que meses son los que faltan.

Vemos en la Figura 5.7 como estarían representados los valores NA.
Se puede observar que hay un alto número de un valor nulo por estación, magnitud al mes, pero esto decae altamente en el resto de valores, teniendo un pequeño repunte en tres valores nulos por estación, magnitud al mes.

Lo que analizamos de la Figura 4.1, es que prácticamente todas las estaciones carecen de alguna medida, salvo siete estaciones. Además, es significativo que, si una medida está a medias, el resto de la misma estación también lo están, siendo los meses de fallo los mismos en todas las magnitudes.

Estos meses no aparecen en los archivos proporcionados por el ayuntamiento, al estar a medias, hemos decidido incluirlos como valores nulos. Esto lo realizamos para preservar el formato, todos los meses deben estar completos para un tratamiento uniforme.

Hemos generado un nuevo archivo, *"meteoDiario21_RellenadoNA.csv"*, con todos estos valores nulos que antes no aparecían.

La red de estaciones cubre todo el municipio de Madrid y obtiene medidas representativas de su entorno, desde el consistorio decidieron que serían estas siete estaciones las que tendrían instaladas todos los sensores: Juan Carlos I, Plaza Elíptica, Ensanche de Vallecas, Casa de Campo, Peñagrande, J.M.D Hortaleza, Centro Mpal. De Acústica, J.M.D. Villaverde, J.M.D. Moratalaz.

Justo en estas estaciones es donde se concentra más masa arbórea, lo que favorece a que estos datos sean menos influenciados por los gases contaminantes de la ciudad, el efecto cañón creado por edificios, etc. En el siguiente enlace del portal del ayun-

	Estacion	MESES QUE FALLAN (A MEDIAS)
J.M.D. Moratalaz	102	
J.M.D. Villaverde	103	
E.D.A.R. La China	104	
Centro Mpal. De Acústica	106	
J.M.D. Hortaleza	107	
Peñagrande	108	6, 7 y 8
J.M.D.Chamberí	109	
J.M.D.Centro	110	
J.M.D.Chamartín	111	7,8,9,10,11 y 12
J.M.D.Vallecas 1	112	
J.M.D.Vallecas 2	113	
Matadero 01	114	4,5,6,7,8,9 y 10
Matadero 02	115	10
Plaza España	4	11 y 12
Escuelas Aguirre	8	
Arturo Soria	16	
Farolillo	18	
Casa de Campo	24	2
Plaza del Carmen	35	
Moratalaz	36	
Cuatro Caminos	38	
Barrio del Pilar	39	
Ensanche de Vallecas	54	
Plaza Elíptica	56	
El Pardo	58	
Juan Carlos I	59	

Figura 5.6: Valores nulos Meteo

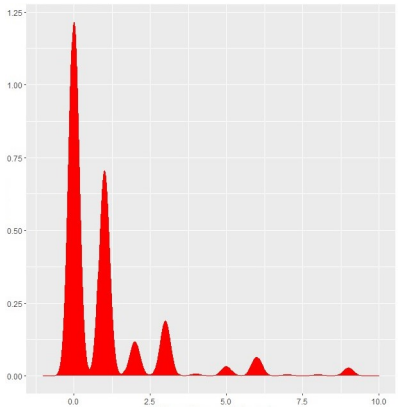


Figura 5.7: Distribución valores NA datos meteorológicos

tamiento[1] podemos corroborar todo lo mencionado en este punto donde también indica los sensores de cada una de las estaciones meteorológicas del municipio.

[1]Red de meteorología

En el mismo script de limpieza creamos un nuevo campo que contabilizaba los datos nulos por fila (contInv), es decir por la tupla formada por estación, magnitud y mes. Estos datos nulos se pueden dar por diversas razones ya que en algunos casos muestran valores, pero su campo de validez está a nulo. Pueden ser desde el fallo de los sensores, mantenimiento de estos, o la medición anómala por parte de los sensores.

Este análisis da como resultado la siguiente Figura 5.8:

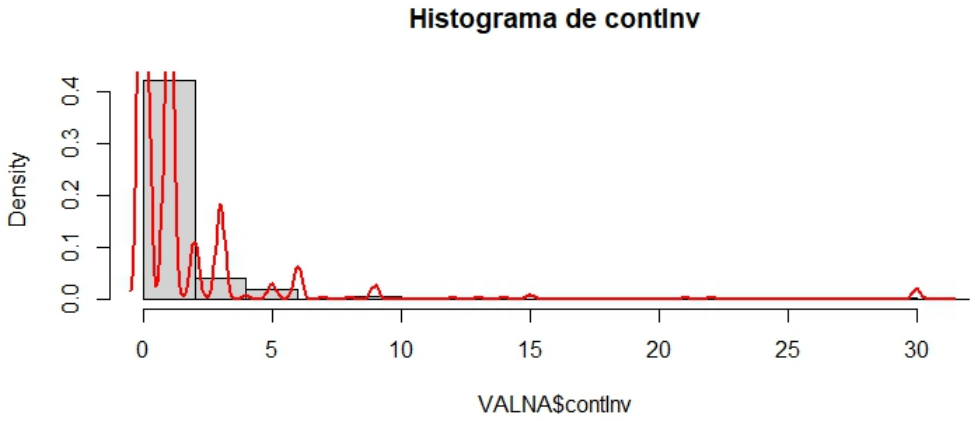


Figura 5.8: Contabilización de los datos nulos por fila en datos meteorológicos

Esto nos muestra que la mayoría de los datos limpiados tienen 0 días con valores inválidos. Algunos meses y algunas mediciones suelen tener entre uno y tres fallos, lo que se puede corresponder fácilmente a un posible mal funcionamiento o fallo y a su correspondiente mantenimiento o reparación. Aunque el promedio está por debajo de 1,5 fallos al mes por sensor. Como se ve claramente en el gráfico presentado, estos fallos siguen una distribución exponencial.

Cabe destacar que hay casos en los que los fallos mensuales llegan a la totalidad del mes, lo que indica un fallo completo del sensor en un mes. Esto ocurre en pocas ocasiones ya que están por debajo de los diez casos de las más de mil mediciones que se realizaron en ese año, lo que supone menos de un 1 % de fallo completo del sensor.

5.3. Valores nulos en datos de contaminación

Al igual que en el punto anterior se realiza una tabla con las magnitudes de las que disponían cada una de las estaciones, Figura 4.4.

También hemos creado la siguiente tabla donde vemos los meses que están a medias y que meses son los que faltan, Figura5.9.

Al igual que en la sección anterior, se observa que todas las estaciones carecen de alguna medida excepto la estación número ocho.

Vemos en la Figura 5.10 que la gráfica es muy parecida a la de los datos meteorológicos, algo que tiene sentido por la explicación dada en la sección 4.2 de *Limpieza*

	Estacion	MESES QUE FALLAN (A MEDIAS)
PZA. DE ESPAÑA	4	11 y 12
ESCUELAS AGUIRRE	8	
AV. RAMÓN Y CAJAL	11	
ARTURO SORIA	16	
VILLAVERDE ALTO	17	2-12
C/ FAROLILLO	18	
CASA DE CAMPO	24	2-12
BARAJAS	27	
PLAZA DEL CARMEN	35	
MORATALAZ	36	
CUATRO CAMINO	38	2-12
BARRIO DEL PILAR	39	
PUENTE DE VALLECAS	40	2-12
MENDEZ ALVARO	47	
CASTELLANA	48	
RETIRO	49	
PLAZA CASTILLA-CANAL	50	
ENSANCHE DE VALLECAS	54	
URBANIZACION EMBAJADA	55	
PLAZA ELIPTICA	56	
SANCHINARRO	57	
EL PARDO	58	
JUAN CARLOS I	59	
TRES OLIVOS	60	

Figura 5.9: Valores nulos contaminación.

de datos donde vemos los fallos o ausencia de sensores y por tanto los valores "NA"de las distintas estaciones.

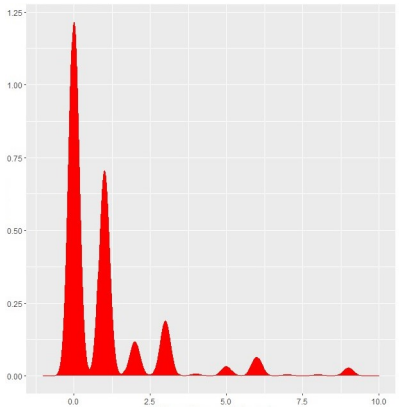


Figura 5.10: Distribución datos contaminación

En el mismo script de limpieza, al igual que en meteo, creamos un nuevo campo que contabilizaba los datos nulos por fila (contInv). Al igual que los datos de meteo,

estos datos nulos se pueden dar por diversas razones.

Este análisis da como resultado la siguiente Figura 5.11:

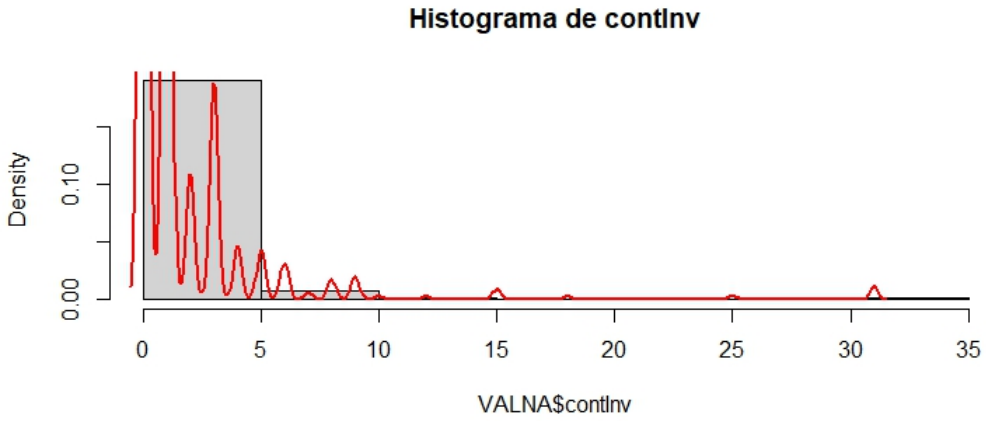


Figura 5.11: Contabilización de los datos nulos por fila en datos contaminación

Al igual que en el histograma de meteo la mayoría de los datos limpiados tienen 0 días con valores inválidos. Algunos meses y algunas mediciones suelen tener entre uno y tres fallos, lo que se puede corresponder fácilmente a un posible mal funcionamiento o fallo y a su correspondiente mantenimiento o reparación. Aunque el promedio está por debajo de 1,5 fallos al mes por sensor.

Los dos histogramas tienen una representación muy parecida.

5.4. Outliers

En este punto detectamos los valores atípicos para analizar y eliminar si fuera necesario del conjunto de datos para que no afecten en los resultados estadísticos y de nuestro estudio.

Para esto vamos a realizar un diagrama de cajas y bigotes, que es una manera muy visual de agrupar datos numéricos a través de sus cuartiles.

Las líneas que se extienden paralelas a las cajas se conocen como «bigotes», y se usan para indicar variabilidad fuera de los cuartiles superior e inferior. Los valores atípicos se representan como puntos individuales que están en línea de los bigotes. Los diagramas de cajas y bigotes se pueden dibujar vertical u horizontalmente.

Normalmente utilizado en estadísticas descriptivas, los gráficos de cajas y bigotes son una excelente forma de examinar rápidamente uno o más conjuntos de datos gráficamente. Estos tienen la ventaja de ocupar menos espacio, lo cual es útil cuando se comparan distribuciones entre muchos grupos o conjuntos de datos.

5.4.1. Datos meteorológicos

En los datos meteorológicos podemos ver primero las gráficas de "bigotes". Todos desplazados hacia abajo excepto en la presión barométrica, ya que sus valores medios

rondan los 950 hPa, Figura 5.12.

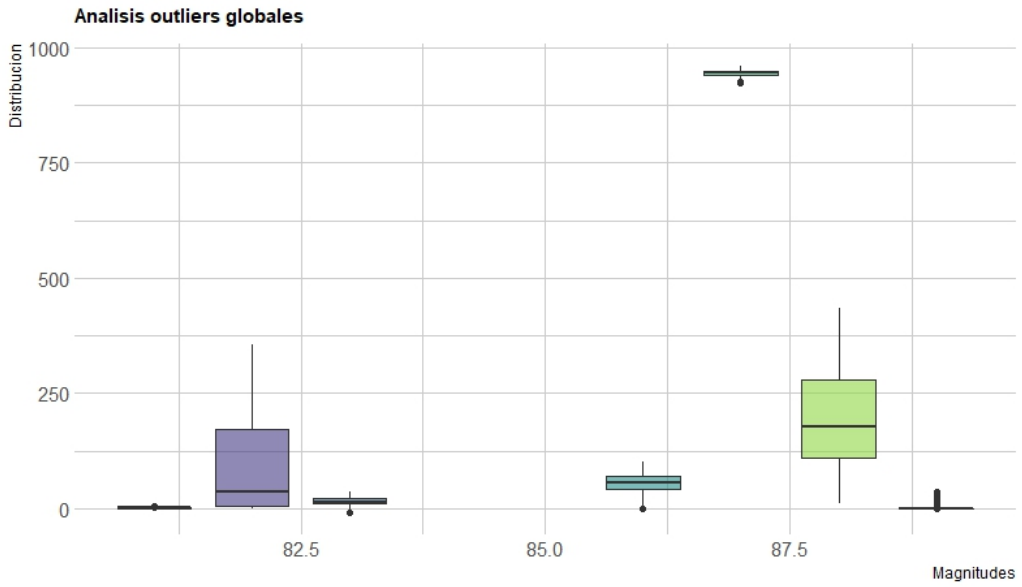


Figura 5.12: Outliner Meteo

Para poder ver si hay alguna anomalía dentro de estos datos pasamos a realizar histogramas individualmente.

En las Figuras 5.13, 5.14, 5.15 la variable velocidad del viento y temperatura obtenemos una función normal, asimétrica y desplazada hacia la izquierda, en la variable Dirección del viento obtenemos una función modal donde la mayor cantidad de datos se encuentran en 0^{o} que equivale a la dirección Norte.

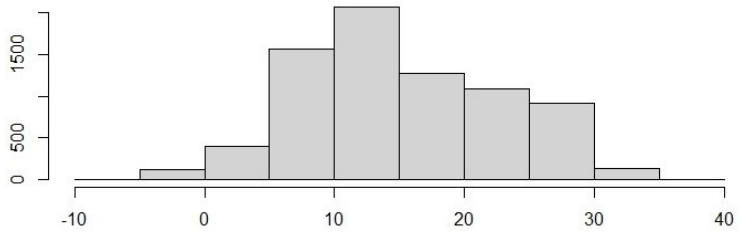


Figura 5.13: Histograma meteo temperaturas

En las Figuras 5.16, 5.17, 5.18 la variable radiación solar obtenemos una función modal y en las variables presión barométrica y humedad relativa tenemos una función normal, ligeramente desplazado a la derecha.

En la Figura 5.19 la variable precipitación tenemos una función normal, asimétrica y desplazada claramente a la izquierda, lo que nos muestra unas lluvias con poca cantidad de precipitación.

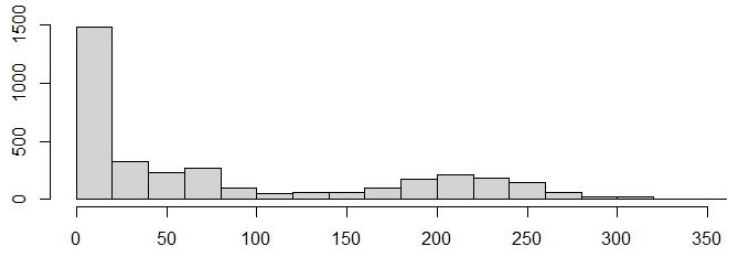


Figura 5.14: Histograma meteo dirección del viento

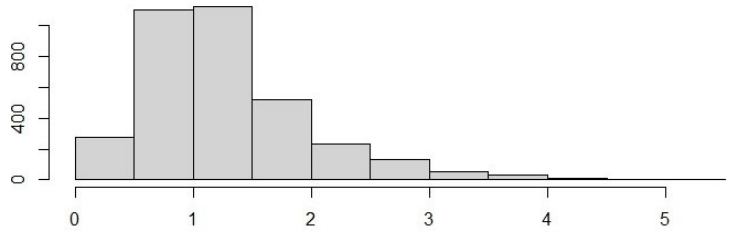


Figura 5.15: Histograma meteo velocidad del viento

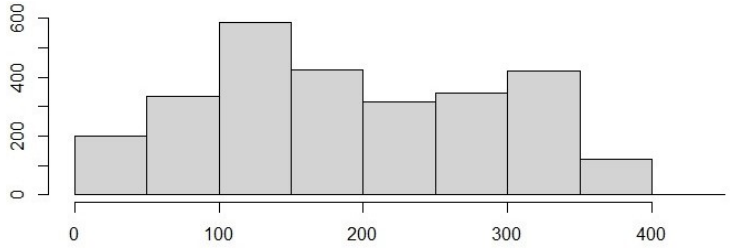


Figura 5.16: Histograma meteo radiación solar

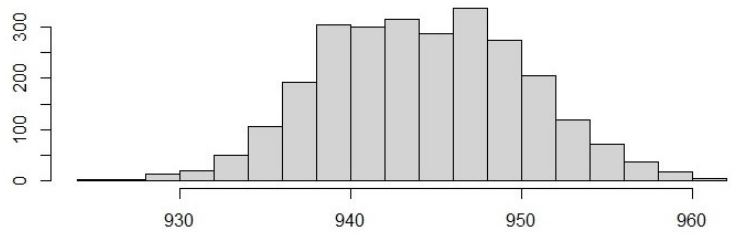


Figura 5.17: Histograma meteo barométrica

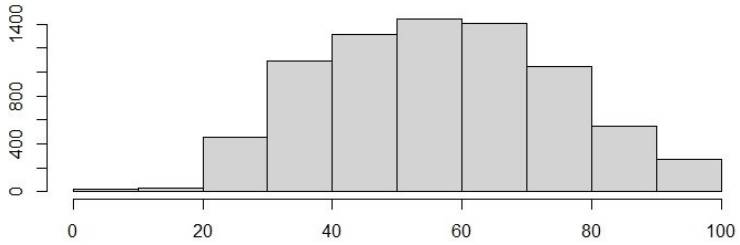


Figura 5.18: Histograma meteo humedad relativa

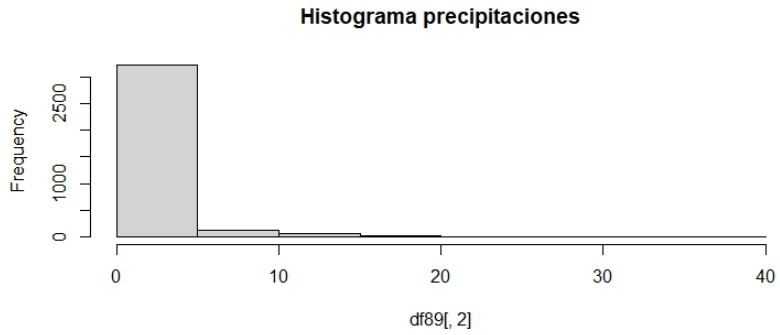


Figura 5.19: Histograma meteo precipitación

5.4.2. Datos contaminación

En los datos de calidad del aire podemos observar que hay valores más atípicos, pero tenemos que tener cuidado ya que pueden ser debidos a acontecimientos meteorológicos o situaciones anómalas que se hayan podido producir. Figura 5.20

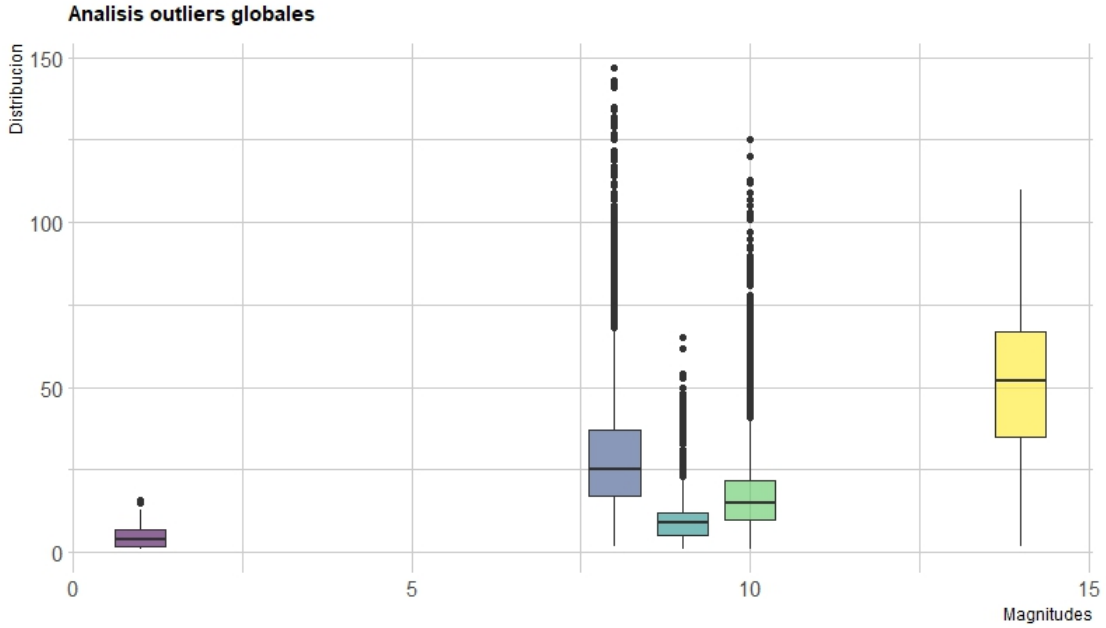


Figura 5.20: Outliner contaminación

Para comprobar lo anterior pasamos a realizar histogramas individuales, Figuras 5.21, 5.22, 5.23, 5.24, 5.25 y observamos que todas las gráficas tienen una función normal desplazada hacia la izquierda excepto en el Ozono, que tiene una función normal, pero con un desplazamiento de los datos más en él, debido a que este contaminante aumenta junto a la temperatura.

Después de analizar las gráficas hemos decidido no eliminar ningún valor ni de

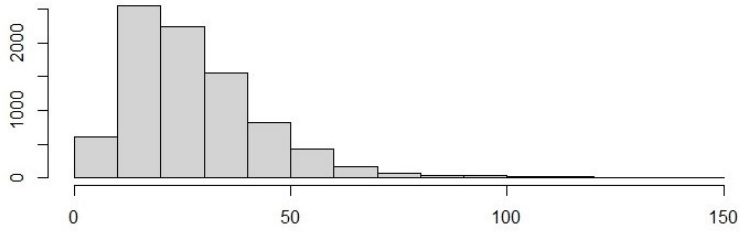


Figura 5.21: Histograma contaminación dióxido de nitrógeno

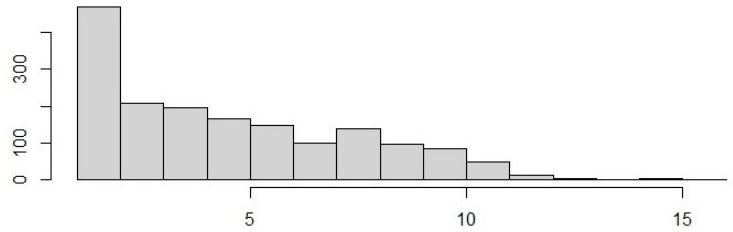


Figura 5.22: Histograma contaminación dióxido de azufre

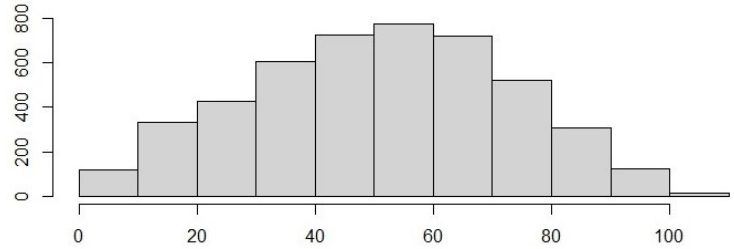


Figura 5.23: Histograma contaminación Ozono

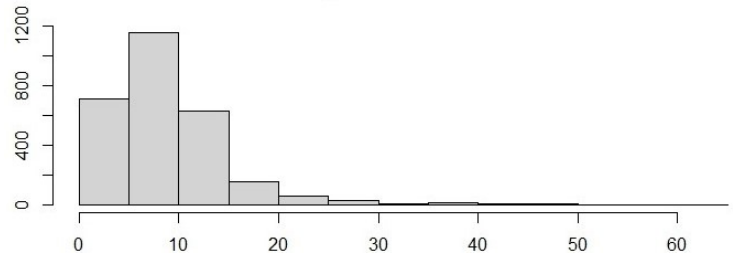


Figura 5.24: Histograma contaminación PM 2.5

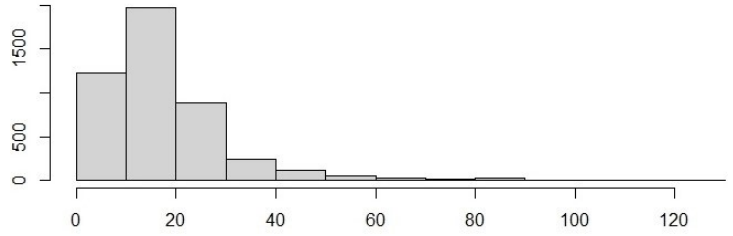


Figura 5.25: Histograma contaminación PM 10

los datos de calidad del aire, ni de los datos meteorológicos, ya que no son factor a

la hora de nuestro análisis.

Capítulo 6

Análisis de datos y extracción de información

En este capítulo, examinamos las relaciones que existen entre los elementos que componen el conjunto de datos. Además, describimos los algoritmos utilizados de Machine Learning para su clasificación. Al completar este capítulo habremos logrado las metas propuestas.

6.1. Clustering

El clustering, es la técnica de Machine Learning que vamos a utilizar para hacer una clasificación no supervisada que consiste en agrupar los datos en diferentes conjuntos homogéneos en función de la similitud de sus características o propiedades.

Vamos a realizar distintas aproximaciones con los datos tratados obtenido del Ayuntamiento de Madrid. Para esto, se ha utilizado los siguientes métodos, el método *K-Means* y el método *Jerárquico*.

6.2. Selección de la estación de calidad del aire y meteorología y tráfico.

Debido a la cantidad de datos vacíos que se nos quedaban y que antiguamente había estaciones que solo median uno de los dos parámetros (meteorológicos o de calidad del aire), hemos eliminado esas estaciones ya que no nos proporcionaban una cantidad de datos con la que poder hacer el estudio, esto se desarrolla en el código a la hora de realizar el *clustering*, ya que dependiendo el tipo de aproximación que realizamos, nos quedamos con unas u otras estaciones.

Para las siguientes aproximaciones vamos a utilizar:

- En la primera aproximación nos quedamos con las siguientes estaciones meteorológicas: 24, 59 ,102, 103, 106, 107. Por tener datos de las siete magnitudes antes mencionadas.

- En la segunda aproximación nos quedamos con las siguientes estaciones meteorológicas: 8, 24, 35, 36, 38, 54, 56, 58, 59, 102, 103, 106, 107, 109, 110, 112. Porque todas ellas tienen las magnitudes 83 y 86.

- En la tercera aproximación nos quedamos con las siguientes estaciones meteorológicas: 8, 18, 24, 36, 38, 54, 58, 59, 102, 103, 106, 107. Por disponer todas ellas de la magnitud 83 y tener datos sobre masas arbóreas.

- En la cuarta aproximación nos quedamos con todas las estaciones de calidad del aire.

- En la quinta aproximación nos quedamos con las siguientes estaciones de calidad del aire: 8, 24, 38, 47, 48, 50, 56. Por tener las magnitudes 8, 9 y 10.

6.2.1. Diagrama del codo

Para elegir el número óptimo de clusters en una aproximación recurrimos a este método.

Este procedimiento utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de Clusters, representando en una gráfica lineal la inercia respecto del número de Clusters. El punto en el que hay un cambio brusco entre las inercias es lo que se llama codo, y ese valor es el que nos muestra el valor recomendable. Este codo suele devolver valores coherentes con respecto al número de clusters de un conjunto de datos.

6.2.2. K-Means

Este algoritmo usa un proceso iterativo para conseguir los grupos, donde se van ajustando hasta obtener el número de agrupaciones convenientes. La ventaja de este algoritmo es que es rápido y sencillo, pero sensible a los outliers.

Vamos a seguir los mismos pasos para las diferentes aproximaciones.

1. Primero aplicamos el algoritmo de *K-Means* y obtenemos el siguiente resultado para la primera aproximación, Figura 6.1:

   Para representarlo de forma que se puedan ver los resultados más claramente, mostramos los datos sobre un mapa como la siguiente imagen de la Figura 6.2.

   Como conclusión de esta primera aproximación podemos observar cómo se crean cuatro grupos, entre ellos tienen características muy parecidas por su ubicación, el Grupo 1 estaría en el centro de una zona residencial y comercial con tráfico (Moratalaz), el Grupo 2 también pero se encuentran más zonas verdes cerca de las estaciones y ubicadas en Madrid-este (Juan Carlos I y Hortaleza), el Grupo 3 se encuentro al lado de casa de campo y en una zona muy abierta pero junto a la M-30 y el Grupo 4 se encuentra fuera del núcleo de Madrid y con parques cerca.

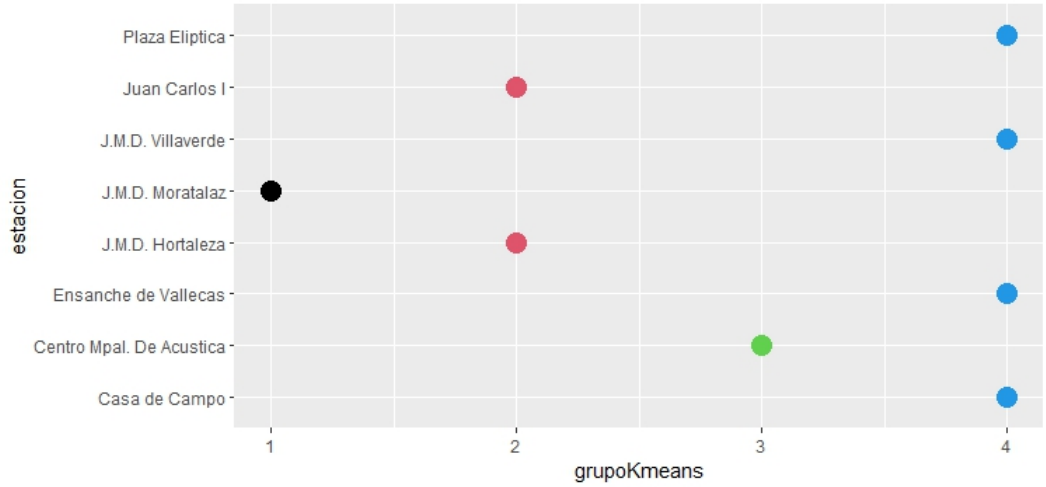


Figura 6.1: K Means primera aproximación

2. Seguimos con la segunda aproximación, en este caso vamos a acotar a las estaciones que contienen las magnitudes 83 (temperatura) y 86 (humedad relativa).

   Primero vamos a obtener el numero óptimo de clusters mediante el **diagrama del codo**. Vemos que la recomendación es de cinco clusters Figura 6.3.

   Aplicamos el algoritmo *K-Means* y obtenemos la Figura 6.4.

   Lo representamos de forma visual y obtenemos la Figura 6.5

   Como conclusión de esta segunda aproximación observamos que salen cinco grupos, el Grupo 1 serían estaciones que se encuentran en parques de grandes extensiones, el Grupo 2 estaciones con parques cercanos pero también cerca de la M-30 o la M-40, el Grupo 3 se encuentra cerca de parques y de zonas abiertas, el Grupo 4 serían estaciones situadas en zonas muy urbanas del centro-sur de Madrid, con poco efecto de zonas verdes a su alrededor y el Grupo 5 seria estaciones en zonas urbanas del centro-norte de Madrid. Por lo que los Grupos 1, 2 y 3 tendrían más zonas verdes y espacios abiertos que los Grupos 4 y 5 quedamos bien reflejados en la Figura 6.5.

3. Continuamos con la tercera aproximación donde vamos a tener en cuenta las zonas verdes de cada una de las estaciones/distritos y estaciones con la magnitud 83 (temperatura).

   Obtenemos primero el numero óptimo de clusters con la normal del codo Figura 6.6.

   Por la imagen vemos que sería cuatro el número aconsejable.

   Pasamos el algoritmo *K-Means* y obtenemos la siguiente Figura 6.7.

   Lo representamos de forma visual obtenemos la siguiente Figura 6.8

   Como conclusión de esta tercera aproximación observamos que salen cuatro grupos, el Grupo 1 serían estaciones que se encuentran en parques de extensión mediana o grande, el Grupo 2 estaciones con parques de mediana extensión, pero cercanos a zonas residenciales y comerciales, el Grupo 3 se encuentra

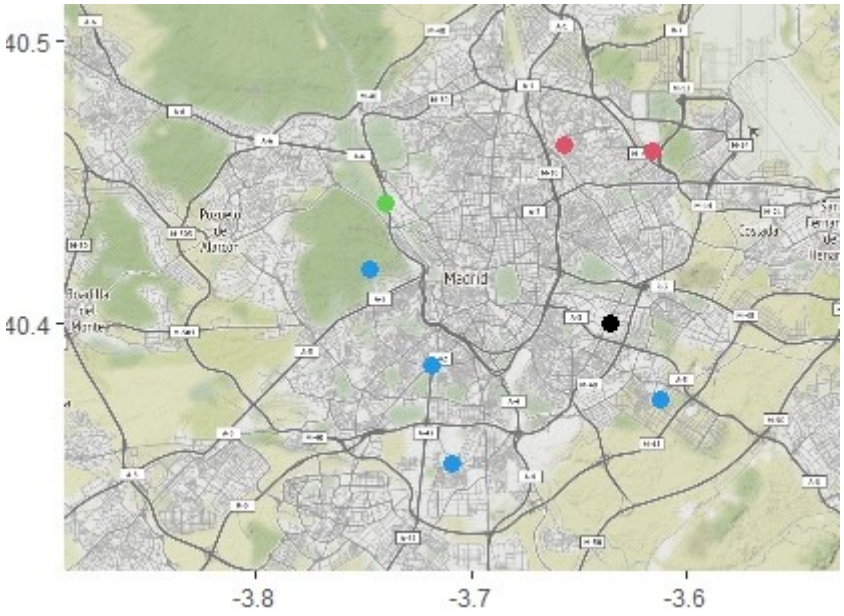


Figura 6.2: Mapa k-Means primera aproximación

cerca de parques y de zonas abiertas y el Grupo 4 serían estaciones situadas en zonas muy urbanas con poco efecto de zonas verdes a su alrededor.

Podemos observar una cierta tendencia hacia el noroeste con similares temperaturas siendo esta el Grupo 1, luego un segundo grupo (Grupo 2) que iría en línea desde la A-5 a la A-2, otro grupo (Grupo 3) con similares temperaturas estando fuera de la M-30 pero con pocas zonas verdes y mucha edificación y por último un grupo (Grupo 4) con grandes extensiones de zonas verdes, pero también cercanos a carreteras principal o zonas urbanas Figura 6.8.

4. Continuamos con la cuarta aproximación donde vamos a tener en cuenta todas las estaciones de calidad del aire y la magnitud 8 (Dióxido de Nitrógeno NO2).

   Pasamos el algoritmo *K-Means* y obtenemos la siguiente Figura 6.9.

   Lo representamos de forma visual obtenemos la siguiente Figura 6.10

   Como conclusión de esta cuarta aproximación observamos que salen cinco grupos, el Grupo 1 serían estaciones que se encuentran en zona urbana con avenidas con tráfico moderado, el Grupo 2 estaciones en zona urbana en cruces de calles con un tráfico alto, el Grupo 3 zonas verdes con poca densidad urbana y de tráfico, el Grupo 4 serían estaciones situadas en zonas muy urbanas con nudos de tráfico y el Grupo 5 se encuentran en zonas urbanas con pocas zonas verdes y densidad de tráfico moderada-alta.

   Podemos observar las zonas que se encuentran cerca de zonas verdes donde los valores de Dióxido de Nitrógeno son más bajos que en otras zonas, Grupo 3, y después tenemos diferentes grupos dependiendo de las zonas verdes que haya cerca y la densidad de tráfico que haya siendo las que menos zonas verdes y más tráfico las que peores datos tienen, Grupo 2 y 4 Figura 6.10.

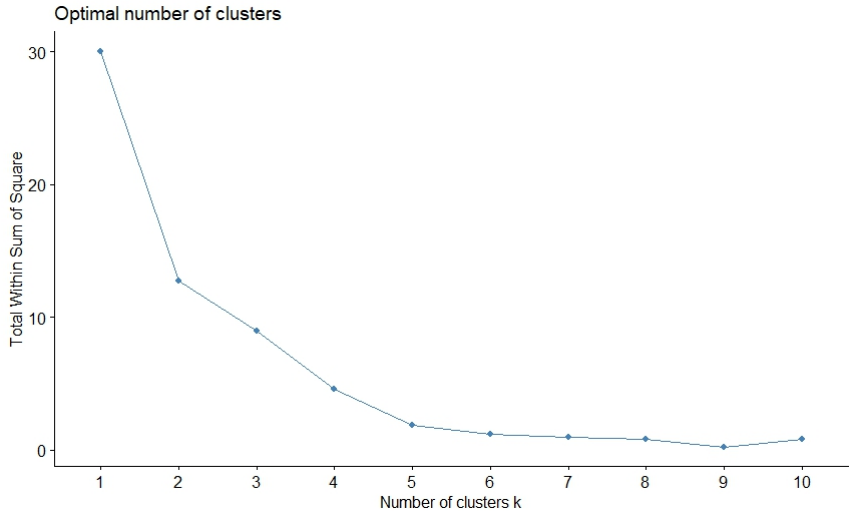


Figura 6.3: Numero óptimo de cluster

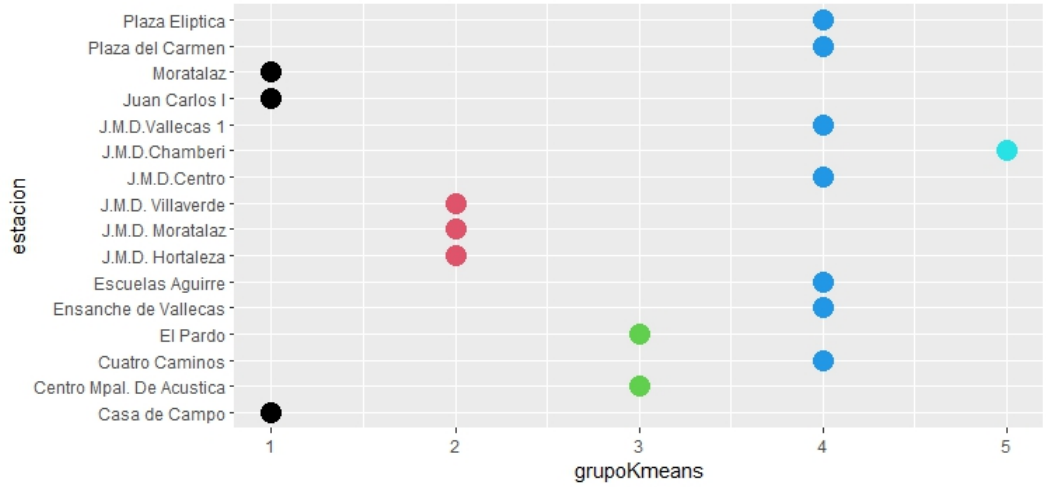


Figura 6.4: K Means segunda aproximación 2

5. Continuamos con la quinta aproximación donde vamos a acotar a las estaciones que contienen la magnitud 8 (Dióxido de Nitrógeno NO2), 9 (Partículas <2.5um) y 10 (Partículas <10um).

   Pasamos el algoritmo *K-Means* y obtenemos la siguiente Figura 6.11.

   Lo representamos de forma visual obtenemos la siguiente Figura 6.12

   Como conclusión de esta quinta aproximación observamos que salen cuatro grupos, el Grupo 1 serían estaciones cerca de grandes avenidas con alta densidad de tráfico, el Grupo 2 estaciones con calles normales con alta densidad de tráfico, el Grupo 3 calles o avenidas grandes con tráfico moderado, pero zonas verdes cerca y el Grupo 4 estaciones en parques con poco o nulo tráfico.

   Podemos observar una tendencia donde en zonas verdes hay datos menos elevados y a la vez que aumentamos zonas con poca masa arbórea y más densidad de tráfico, peores datos contaminantes tenemos 6.12.

Figura 6.5: Mapa K Means segunda aproximación 2

Si observamos la aproximación tres 6.7 y la aproximación cinco 6.11 vemos que se corresponden las estaciones meteorológicas con mayor temperatura con las estaciones de calidad del aire con mayor contaminación. Por lo que podemos corroborar que el efecto de los contaminantes, junto a zonas con poca masa arbórea en zonas urbanas con alta densidad de edificios provoca que el efecto isla de calor aumente. Al igual que se corresponden la estación de Casa de Campo en los dos análisis, esto nos ayuda a ver que en sentido contrario los contaminantes también afectan a la temperatura en las estaciones y que en este caso el estar cerca de una gran masa arbórea ayuda a que las temperaturas sean más suaves y los elementos contaminantes no afecten tanto a esta zona.

6.2.3. Jerárquico

El método jerárquico permite la construcción de un dendrograma, en el cual se puede seguir de forma gráfica el procedimiento de unión, mostrando que grupos se van uniendo, en qué nivel concreto lo hacen, así como el valor de la medida de asociación entre los grupos cuando estos se agrupan (valor que llamaremos nivel de fusión). En resumen, la forma general de operar de estos métodos es bastante simple.

Un dendrograma es un tipo de representación gráfica en forma de árbol que organiza y agrupa los datos en subcategorías según su similitud; dada por alguna medida de distancia. Los objetos similares se representan en el dendrograma por medio de un enlace cuya posición está determinada por el nivel de similitud entre los objetos o grupos de objetos. Dadas estas características, hace que los dendrogramas sean un tipo de diagrama muy útil para estudiar las agrupaciones de objetos; es decir, para estudiar los Clusters que pueden darse en un data-set.

Al igual que en el método *K-Means*, vamos a seguir los mismos pasos.

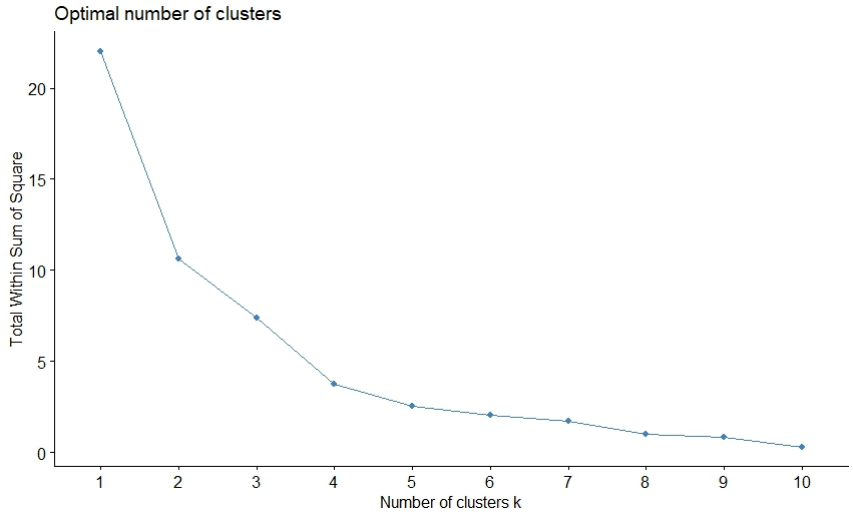


Figura 6.6: Numero de Cluster zonas verdes

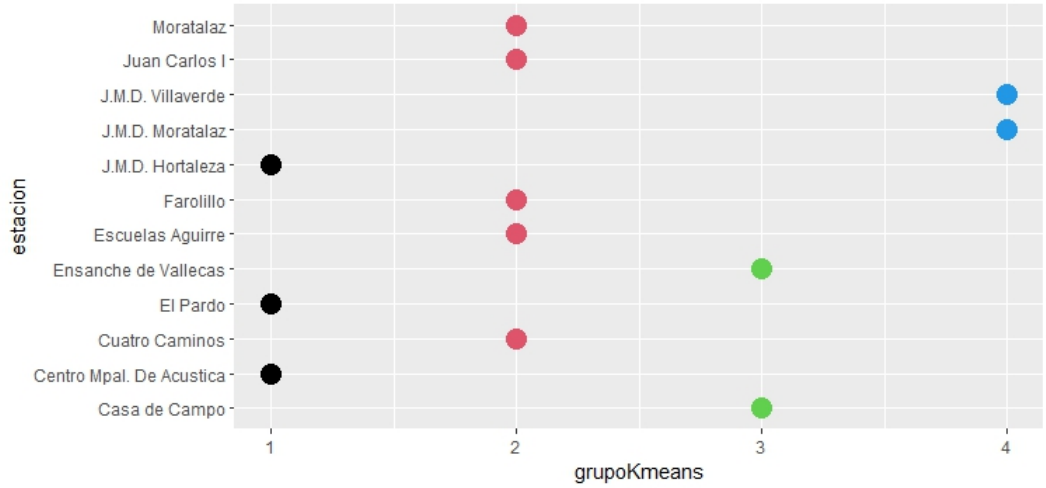


Figura 6.7: K Means tercera aproximación 3

1. Primero pasamos el algoritmo del método Jerárquico ante una primera aproximación y obtenemos el siguiente resultado Figura 6.13.

   Lo representamos de forma visual obtenemos la siguiente Figura 6.14

   Aquí como conclusión podemos establecer la siguiente relación entre las estaciones, en Rojo podemos agrupar las estaciones que se encuentran en Madrid zona sur fuera de la M-40, Verde que sería colindante a la M-30, Azul que sería parque en el exterior con poco efecto de isla de calor y Negro, zonas urbanas con pocas zonas verdes.

2. Seguimos con la segunda aproximación, en este caso vamos a acotar a las estaciones que contienen las magnitudes 83 (temperatura) y 86 (humedad relativa).

   Pasamos el algoritmo *Jerárquico* y obtenemos la Figura 6.15

   Lo representamos de forma visual obtenemos la siguiente Figura 6.16

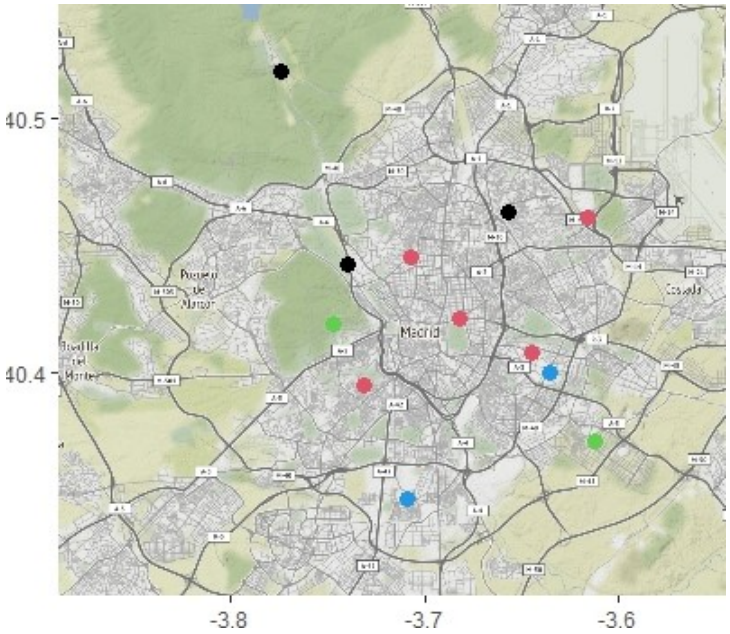


Figura 6.8: Mapa K Means tercera aproximación 3

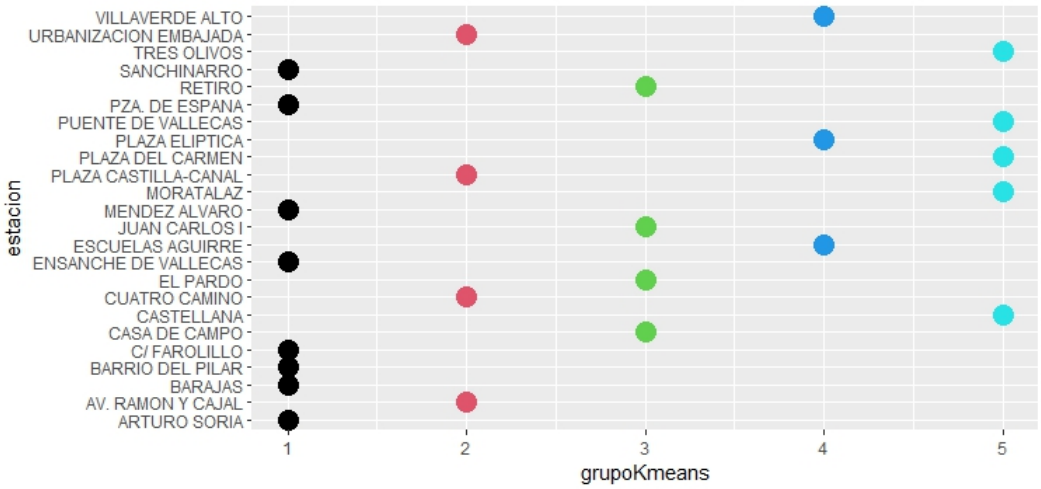


Figura 6.9: K Means cuarta aproximación 4

Como conclusión de esta segunda aproximación observamos que salen cinco grupos, en Rojo tendríamos estaciones que se encuentran en zonas con grandes extensiones de zonas verdes cerca, pero con transito como la M-30, en azul claro tendríamos estaciones con parques de gran extensión cerca, en color Azul oscuro tenemos estaciones que se encuentra cerca de parque, pero en zonas urbanas, en Negro tendríamos zonas urbanas sin parques cerca. Aquí si tenemos más información y podemos ver que hay claras zonas definidas. 6.16.

3. Continuamos con la tercera aproximación donde vamos a tener en cuenta las zonas verdes de cada una de las estaciones/distritos y estaciones con la magnitud 83 (temperatura).

   Pasamos el algoritmo *Jerárquico* y obtenemos la siguiente Figura 6.17

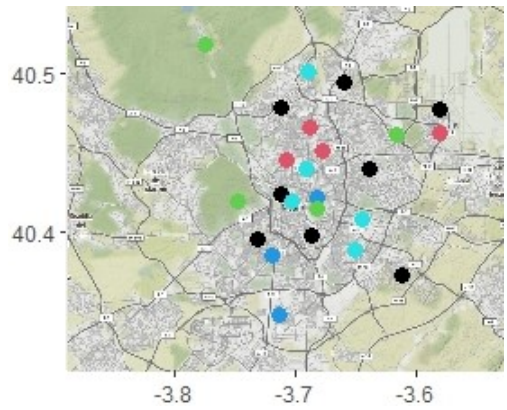


Figura 6.10: Mapa K Means cuarta aproximación 4

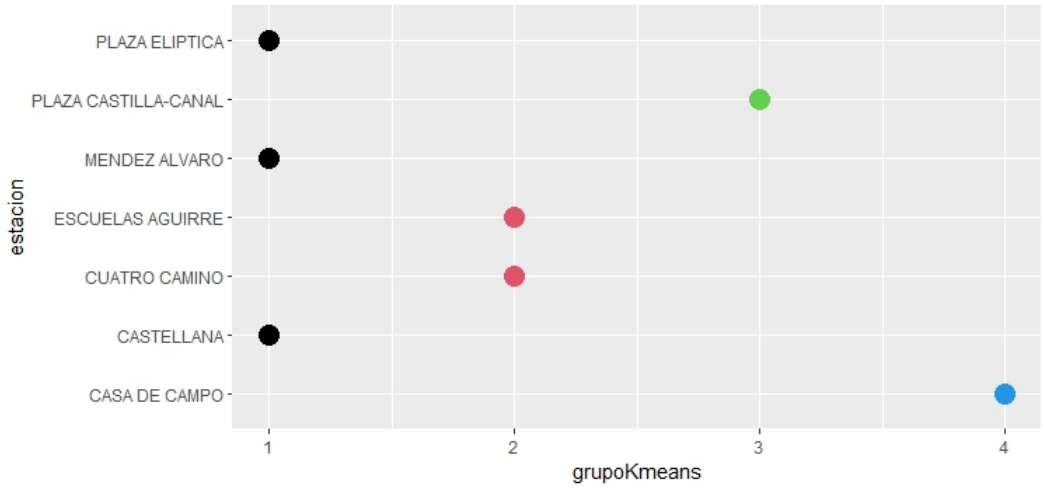


Figura 6.11: K Means quinta aproximación 5

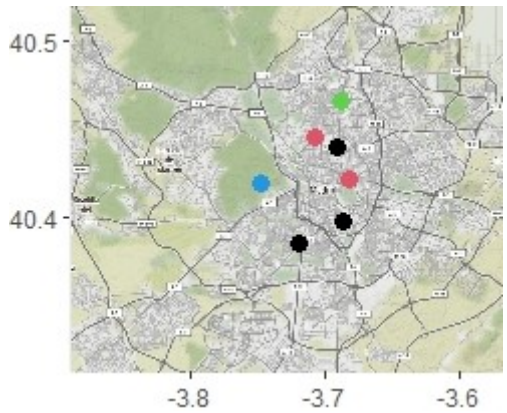


Figura 6.12: Mapa K Means quinta aproximación 5

Lo representamos de forma visual obtenemos la siguiente Figura 6.18

Como conclusión de esta tercera aproximación observamos que salen cuatro grupos, Verde zonas cercanas a grandes masas arbóreas por lo que la temperatura hace que estén en el mismo intervalo, Azul serian estaciones que se

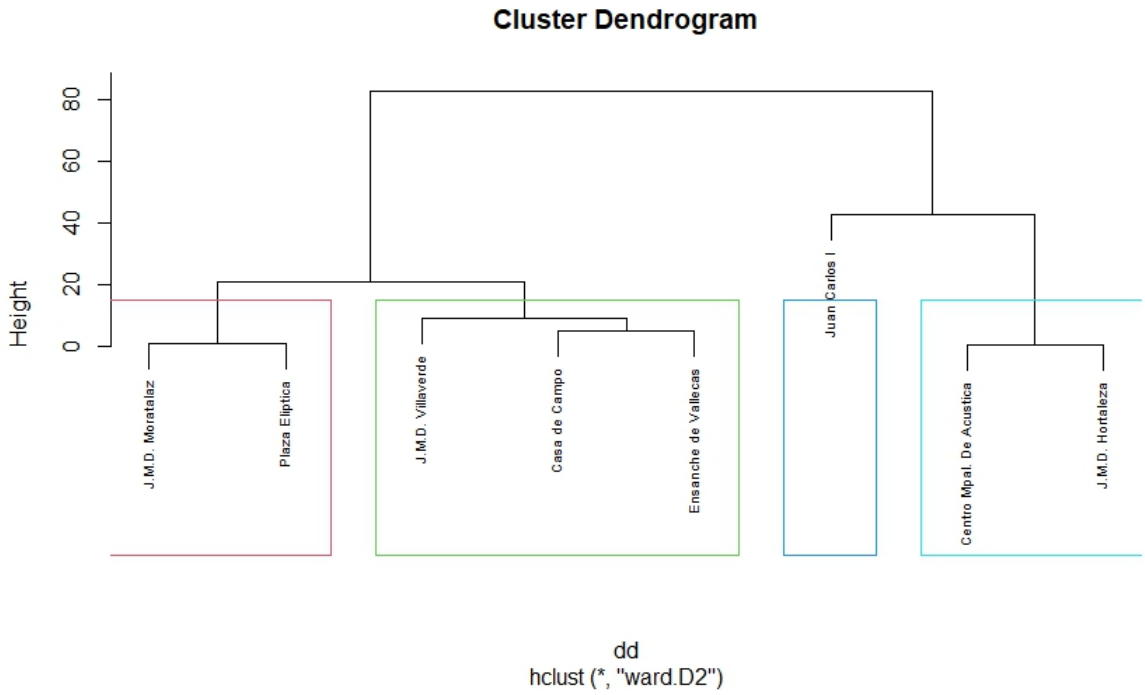


Figura 6.13: Jerárquico Aproximación 1

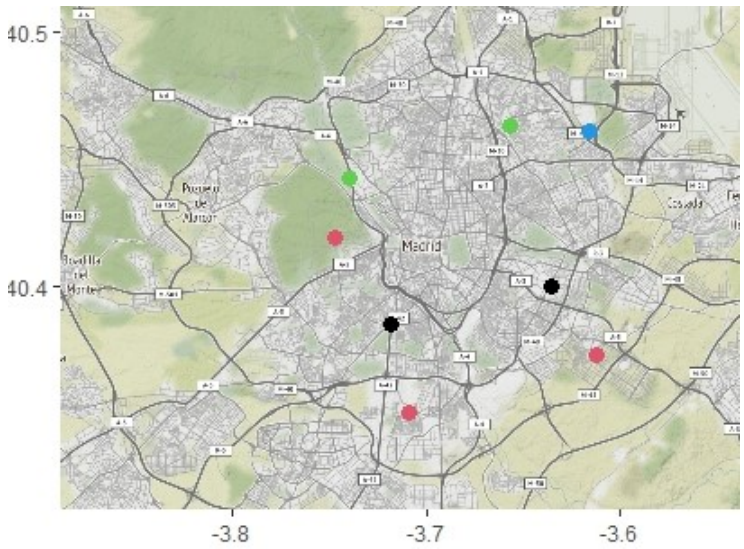


Figura 6.14: Mapa Jerárquico Aproximación 1

encuentran en parques de extensión mediana o grande por lo que la temperatura actúa igual, en Rojo tenemos también similitud estando generalmente más dentro de una zona urbana pero teniendo cerca algún parque grande como puede ser el Juan Carlos I y por ultimo tenemos las estaciones en Negro donde son estaciones que no tienen ningún parque cerca y están en mitad de zonas urbanas densas por lo que se observa que ese efecto *isla de calor* es mayor en zonas que no tienen parques cerca.

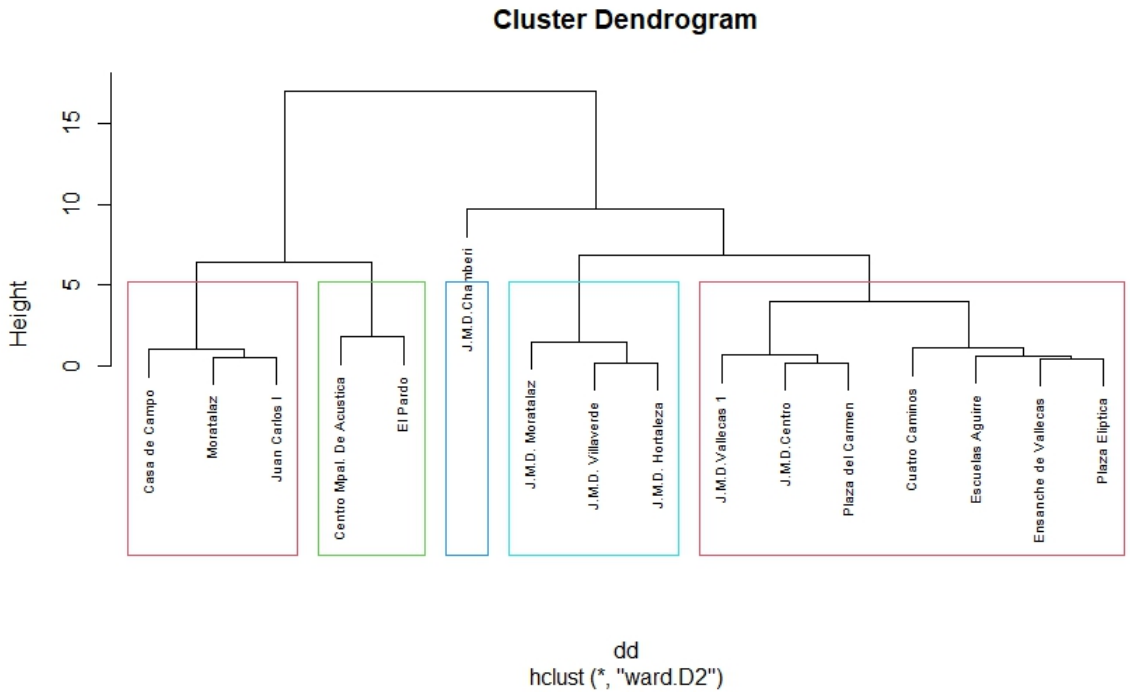


Figura 6.15: Jerárquico Aproximación 2

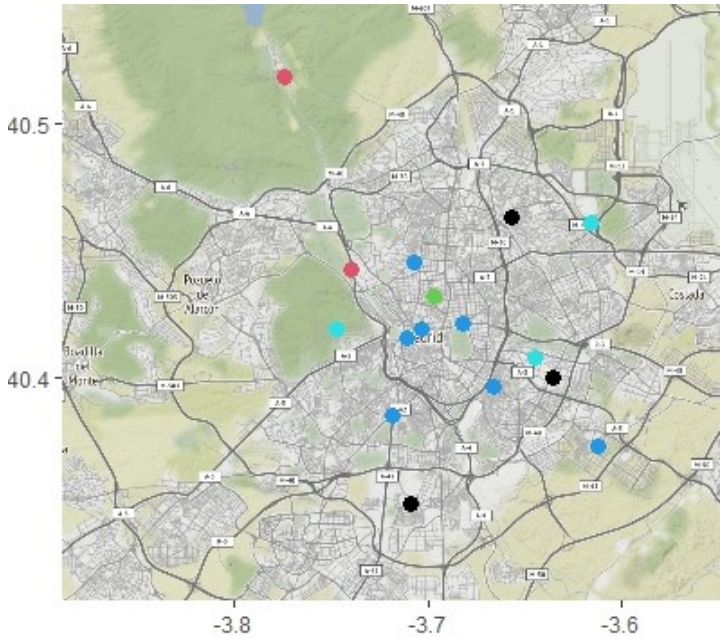


Figura 6.16: Mapa Jerárquico Aproximación 2

4. Continuamos con la cuarta aproximación donde vamos a tener en cuenta todas las estaciones de calidad del aire y la magnitud 8 (Dióxido de Nitrógeno NO2).

   Pasamos el algoritmo *Jerárquico* y obtenemos la siguiente Figura 6.19

   Lo representamos de forma visual obtenemos la siguiente Figura 6.20

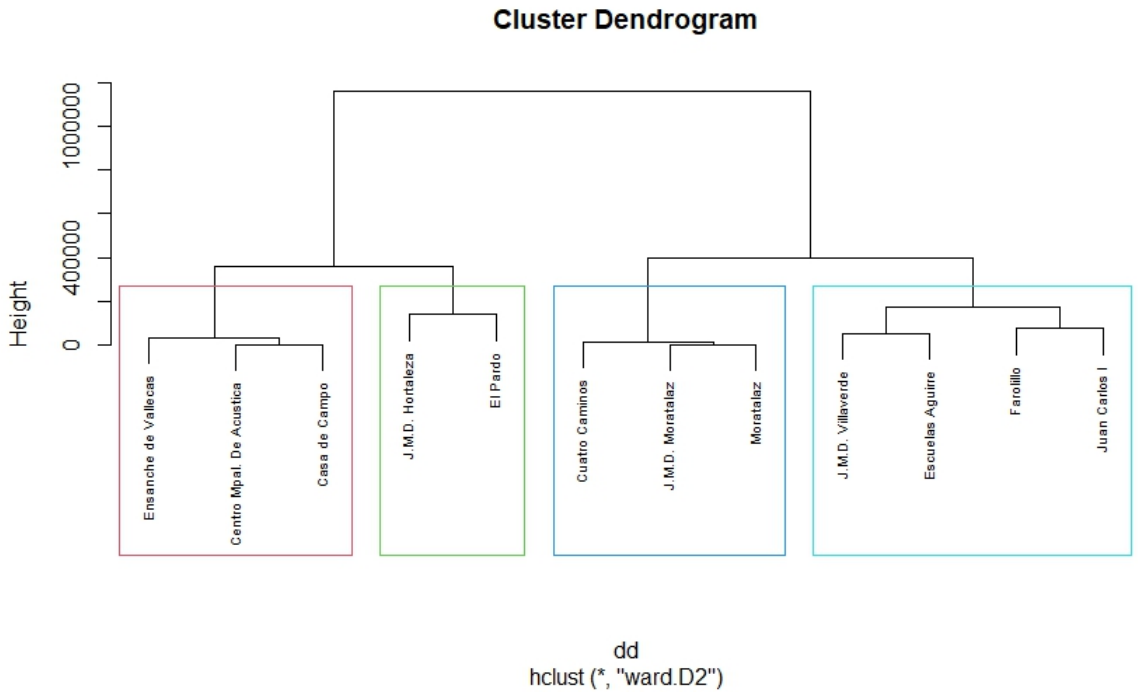


Figura 6.17: Jerárquico Aproximación 3

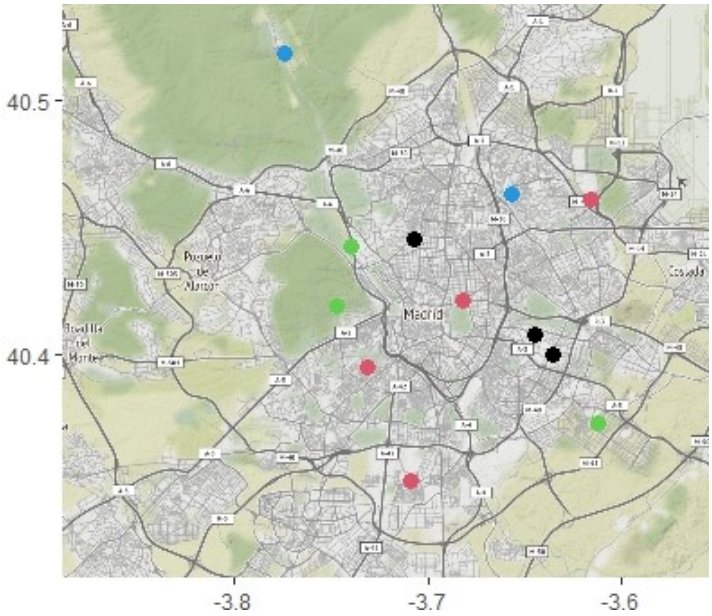


Figura 6.18: Mapa Jerárquico Aproximación 3

Como conclusión de esta cuarta aproximación observamos que salen cinco grupos, en general percibimos cierto comportamiento similar en ellos, pero por la gran cantidad de datos valoraros aquí no sacamos ninguna conclusión clara. En general vemos que hay una similitud en las zonas periféricas, pero hay estaciones muy cercanas en diferentes grupos.

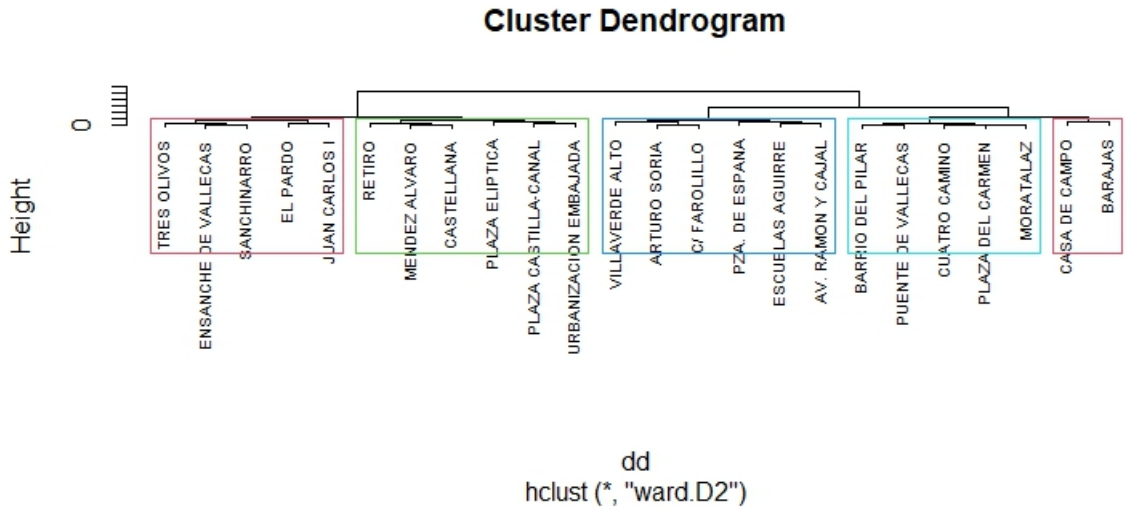


Figura 6.19: Jerárquico Aproximación 4

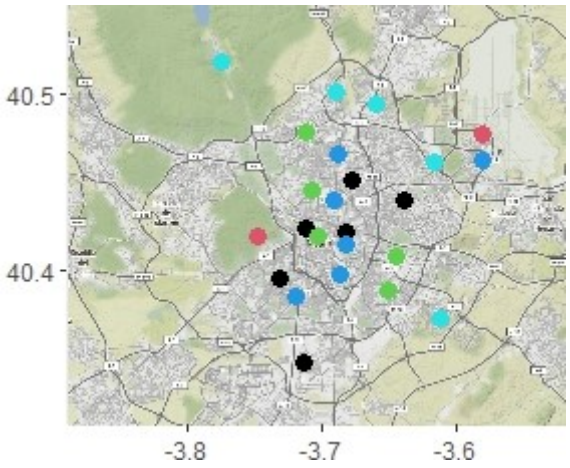


Figura 6.20: Mapa Jerárquico Aproximación 4

5. Continuamos con la quinta aproximación donde vamos a acotar a las estaciones que contienen la magnitud 8 (Dióxido de Nitrógeno NO2), 9 (Partículas <2.5um) y 10 (Partículas <10um).

   Pasamos el algoritmo *Jerárquico* y obtenemos la siguiente Figura 6.21

   Lo representamos de forma visual obtenemos la siguiente Figura 6.22

   Aquí sin embargo si podemos sacar conclusiones, ya que se ve como estos contaminantes tienen una menor presencia en el grupo Rojo y todo lo contrario en el Grupo verde que se encuentra en zonas urbanas y con alta densidad de tráfico, por lo que aumenta los efectos de la *isla de calor*

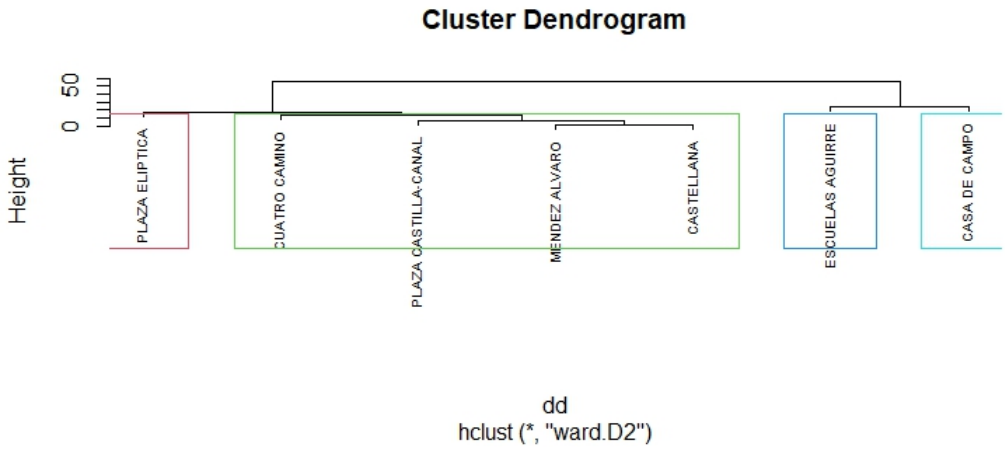


Figura 6.21: Jerárquico Aproximación 5

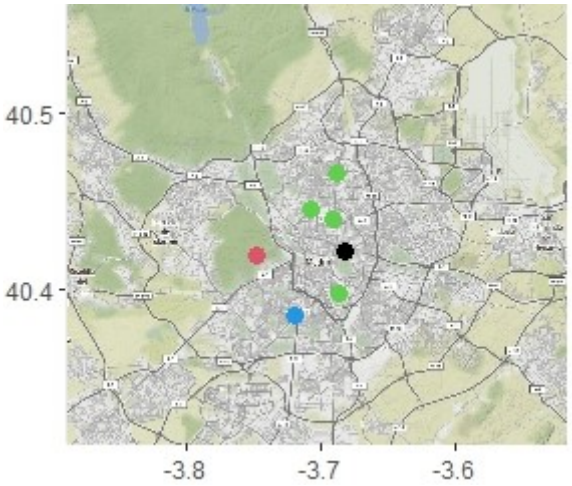


Figura 6.22: Mapa Jerárquico Aproximación 5

6.3. Correlaciones

El propósito de esta sección es identificar las relaciones que existen para las diversas variables relevantes para cada conjunto de datos, así como entre el conjunto completo de variables disponibles para la implementación del proyecto.

Se utiliza el método de correlación canónica en R. Que es una técnica estadística de análisis multivariantes.

Una correlación es una relación proporcional de una agrupación de, principalmente 2, variables numéricas, que al variar una de ellas, la otra lo hace del mismo modo. Esto explica o demuestra que un conjunto de datos sufre una evolución creciente o decreciente si las variables que lo componen están coleccionadas.

Es por esto por lo que vamos a medir la correlación entre las distintas variables que componen nuestro conjunto de datos (contaminantes, meteorológicos y masa arbórea)

Para ello sacaremos la media de las distintas variables a lo largo del año y veremos

cómo se comportan.

Por el tema que tratamos, valoramos que la temperatura es la magnitud más importante a analizar. Por eso empezamos a analizar los datos que hemos limpiado anteriormente y obtenemos el siguiente ejemplo Figura 6.23) donde se ve la evolución de la temperatura en la estación de Moratalaz (numero 102).

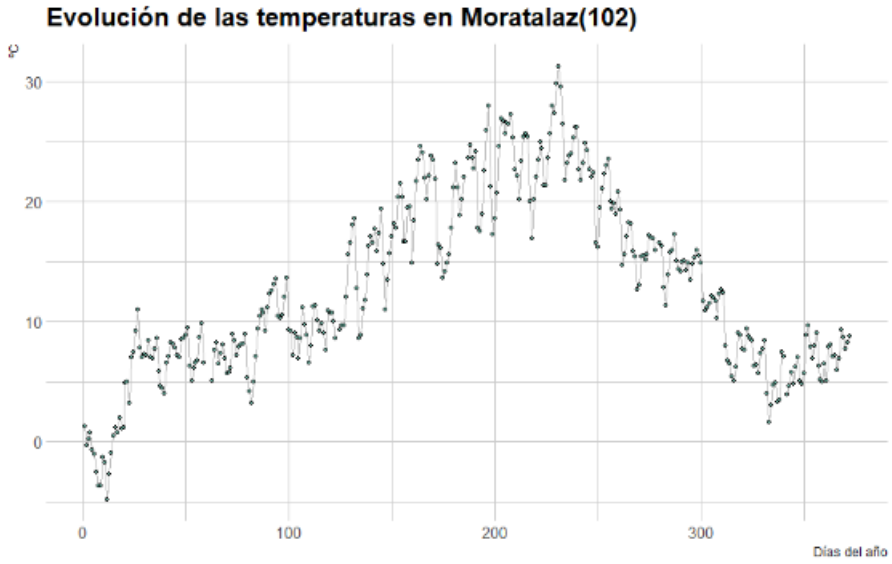


Figura 6.23: Evolución de temperatura Moratalaz

En ella podemos observar valores lógicos, siendo los más cercanos al equinoccio de invierno más frías y acercándose al equinoccio de verano más altas.

En la siguiente Figura podemos observar la correlación entre dos estaciones, la de Moratalaz anteriormente vista y la estación Centro (Numero 110) Figura 6.24

Aquí podemos observar que durante todo el año las temperaturas en la estación Centro, son más altas que en la estación Moratalaz.

En la siguiente figura hemos mezclado todas las medidas de temperatura de todas las estaciones, depurándolas hemos dejado las cinco con temperatura más extremas tanto por frío como por calor, quedando la siguiente Figura 6.25

En el observamos que las estaciones son la 102-Moratalaz, 103-Villaverde, 106-Moncloa/Aravaca, 109-Chamberi y 112-Villa de Vallecas. Siendo las más extremas por arriba Chamberí y Villa de Vallecas y las otras tres, las más extremas por abajo. Se ve como en enero la estación de Moncloa se acerca a los -10º mientras que en Chamberí se queda unos grados por encima de 0º.

Empezamos a hacer correlaciones sencillas, por ejemplo, en la siguiente Figura 6.26, nos reafirmamos en algo lógico como seria pensar que la radiación solar será baja los días nublados o con precipitaciones. Figura 6.26

Vemos como si se corresponde esa correlación, es verdad que hay días donde no hay precipitaciones y baja la radiación siendo días nublados o con alta nubosidad.

Continuamos realizando una correlación temperaturas/humedad relativa Figura 6.27. Como ya habíamos entendido de Tejedor et al. (2016), "*en los periodos cálidos*

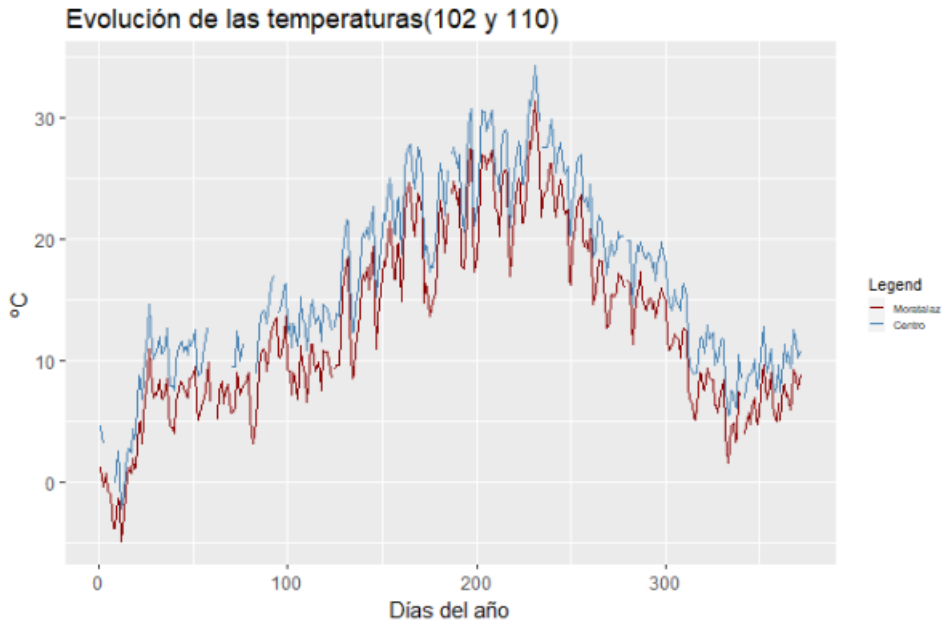


Figura 6.24: Evolución temperatura Moratalaz Centro

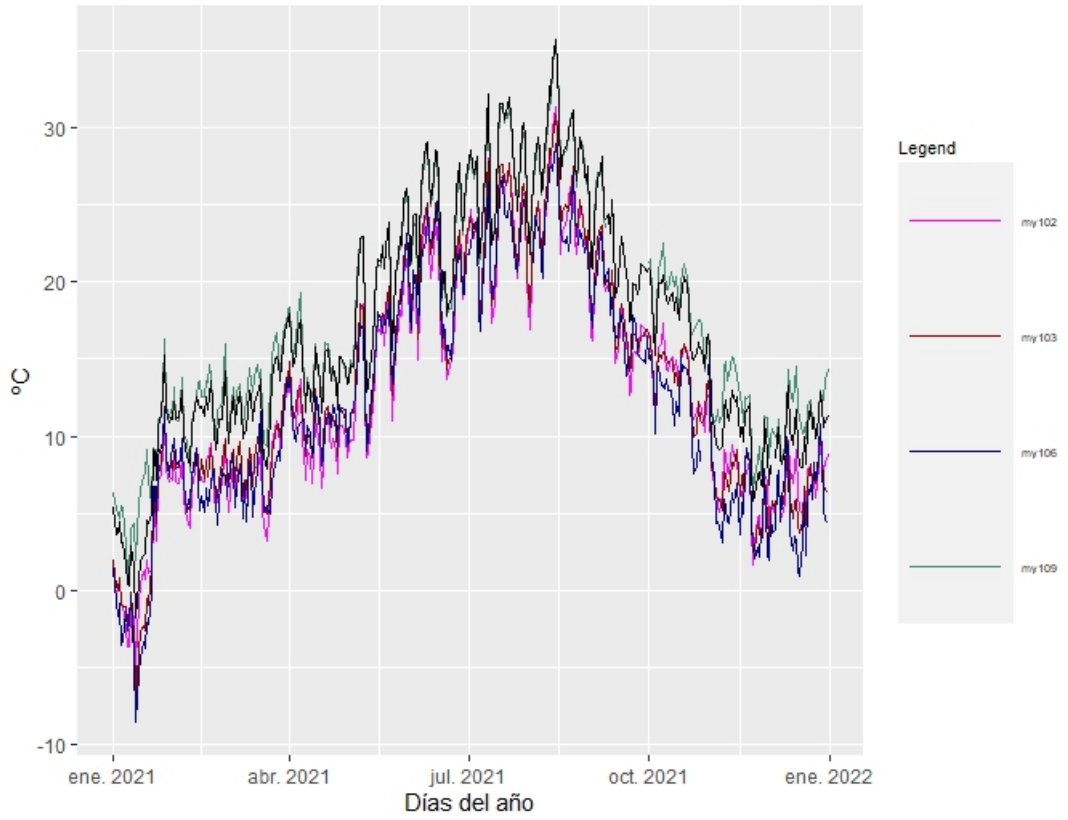


Figura 6.25: Estaciones con temperatura más extrema

la ciudad de Zaragoza agudiza de forma considerable el calor y la sequedad en relación a las zonas no urbanas", y de igual modo pasa en la ciudad de Madrid.

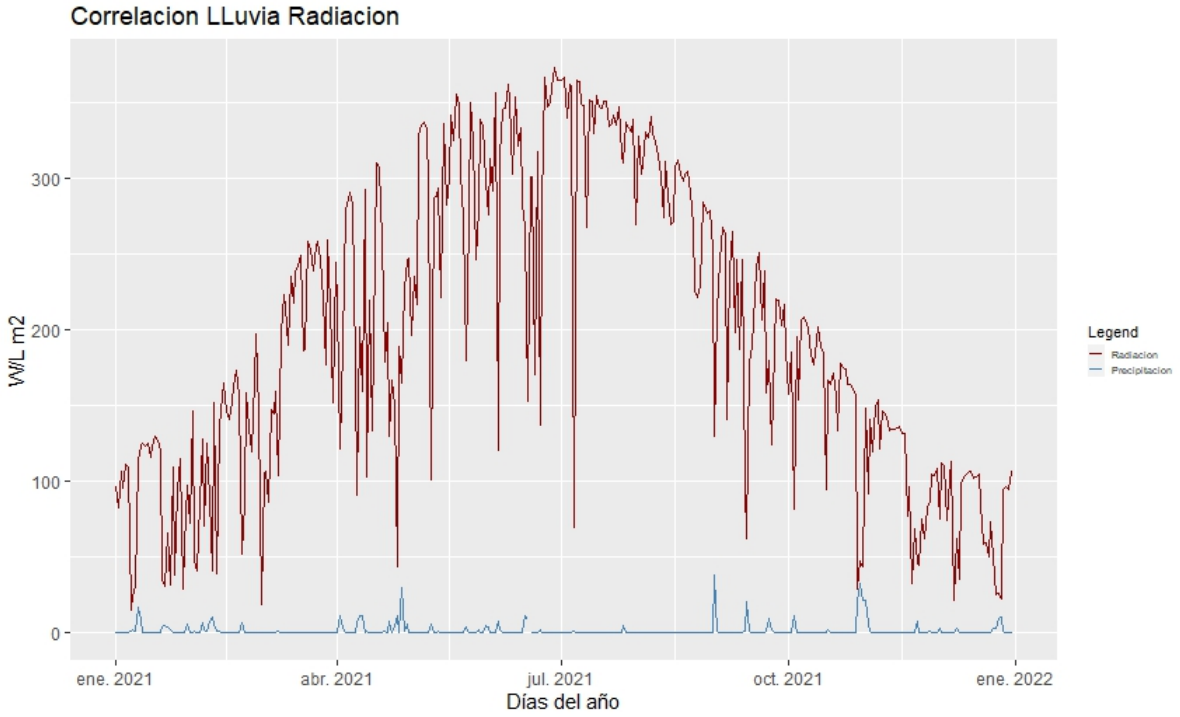


Figura 6.26: Correlación lluvia radiación

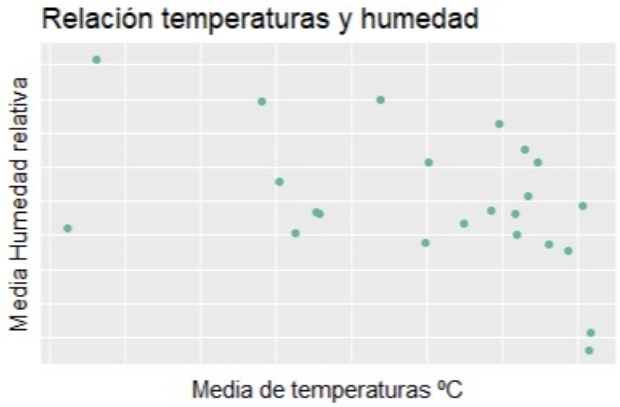


Figura 6.27: Correlación temperatura/humedad

Donde podemos observar que, si hay una relación, siendo las zonas más húmedas las que menor media de temperaturas tienen y viceversa.

Continuamos realizando una correlación masa arbórea/ temperaturas Figura 6.28, dónde cada uno de los puntos representa una estación.

Podemos observar que a mayor cantidad de masa arbórea hay una menor temperatura, por lo que se puede concluir que en las zonas verdes hay una suavización de las temperaturas.

Concluimos con que el efecto ICU tiene una localización central y que se desplaza hacia sur y este. Así mismo existen diferencias de temperatura entre áreas urbanas y

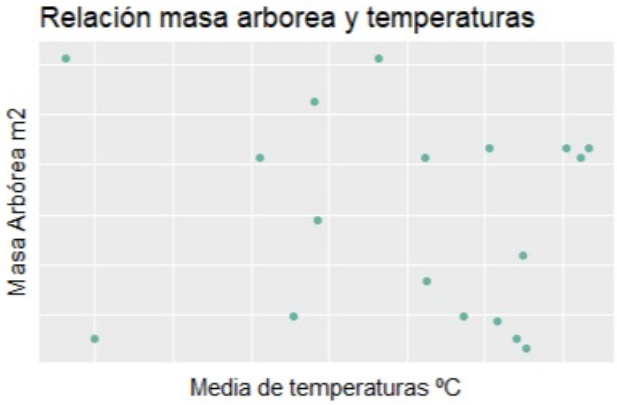


Figura 6.28: Correlación arbórea/temperatura

no urbanas. Y dentro de las zonas urbanas se ven diferenciados grupos entre aquellas que tienen zonas verdes y azules. En el Capítulo 7 detallamos en profundidad el resto de las conclusiones.

Capítulo 7

Conclusiones y Trabajo Futuro

7.1. Conclusiones

En vista de los resultados obtenidos durante la realización de este trabajo y teniendo en cuenta los objetivos definidos al comienzo en la Sección 1.2, hemos obtenido las siguientes conclusiones.

O1. Se ha tenido que realizar el limpiado, formateado y filtrado de datos llegando a recogerlo en un único fichero. Se tuvo que realizar un cambio en el formato que se recoge en el punto 4, que incremento la dificultad a la hora de realizarlo. Quitando esta dificultad se ha conseguido unificar en un único fichero gracias a la herramienta y librerías de R.

Del cumplimiento de este objetivo podemos concluir que el formato proporcionado por el propio Ayuntamiento es mejorable. En él se establecen meses con más días de los que les corresponden, como febrero con 31, lo cual puede llevar a confusiones. Se pueden proponer unos más intuitivos y manejables, como el que hemos realizado en la limpieza, lo que facilita la compresión de estos.

O2. Se han podido aplicar algoritmos y métodos aceptados en el mundo del Data Science para poder realizar distintas clasificaciones de las estaciones. De estas hemos obtenido nuevas conclusiones:

- Las zonas verdes influyen notoriamente en las temperaturas. Agrupando en mismos clusters grandes parques de la ciudad con abundante masa arbórea, así como con presencia de zonas azules como estanques, fuentes o el propio río. Las temperaturas se ven suavizadas en estos puntos de la ciudad, viendo reducido el efecto ICU.

- Los núcleos urbanos se recogen en otro grupo. Se caracterizan por las elevadas temperaturas por encima de las citadas en el punto anterior, así como por ser zonas bastante contaminadas. En estas agrupaciones nos damos cuenta de que el efecto cañón, mencionado en el Estado de la Cuestión, es un hecho probado. Zonas en las que abundan edificios y las velocidades del viento son menores, acumulan calor en el ambiente. Esto influye directamente en las temperaturas y por lo tanto en el efecto ICU.

- Las zonas próximas a las carreteras M-40 y M-30 consolidan otro conjunto. Principalmente debido a los materiales de construcción de estas, que absorben e irradian calor más lentamente. Del mismo modo que el punto anterior son zonas bastante contaminadas, pero que, por lo general, al ser abiertas limitan el efecto cañón.

- Otra de las principales conclusiones es la masificación urbana en Madrid. Únicamente los distritos más exteriores son los que cuentan con grandes masas arbóreas que suavizan las temperaturas. Cuanto más nos acercamos al centro de la ciudad las zonas verdes bajan notoriamente, salvo excepciones como el Parque del Retiro o el Parque del Oeste.

- La presencia de edificios, carreteras y otros elementos urbanísticos, además de limitar el flujo del aire necesario para reducir las temperaturas, se componen de materiales que las aumentan. Estos absorben calor y se irradian lentamente, lo que hace que las diferencias de temperaturas sean más apreciables entre las zonas urbanas y las exteriores con menos edificaciones y zonas verdes.

- Aunque el efecto ICU es notorio durante todo el año, durante las temporadas invernales y estivales se intensifican. Es decir, las temperaturas son más extremas en verano, que sumado al uso de electrodomésticos para paliar las elevadas temperaturas incrementan dicho efecto. Y durante el invierno se observan temperaturas impropias de la estación, pudiendo haber diferencias en torno a los 7-8^{o}C de media entre el centro y las zonas verdes exteriores.

- Teniendo en cuenta las conclusiones anteriores podemos ubicar el efecto ICU en Madrid centro principalmente, donde hay una carencia de zonas verdes, un mayor efecto cañón y unos niveles de contaminación superior al resto de la ciudad. La isla de calor urbana se va disipando hacia el norte, sur y este, que es donde empiezan a aparecer barrios construidos a finales del siglo XX y comienzos del XXI. Sus principales características son los espacios abiertos y grandes avenidas (limitan el efecto cañón), así como la creación de grandes zonas verdes como el Cementerio de la Almudena, la Cuña Verde, el parque Juan Carlos I y el Felipe VI entre otros.

O3. Se ha podido hacer uso de herramientas de visualización de datos para ilustrar estas agrupaciones para una mayor facilidad a la hora de interpretar los resultados, mediante gráficas y mapas siendo estas importante para poder confirmar y observar el efecto *isla de calor*.

O4. Hemos concluido que las carreteras y contaminantes tienen un efecto negativo sobre las temperaturas y calidad del aire, de los que podemos establecer una relación directa. Del mismo modo se aprecia con el exterior de la ciudad y zonas verdes y la mejor calidad.

7.2. Recomendaciones

En vistas a las conclusiones obtenidas sumado al creciente cambio climático, del efecto ICU se espera que siga aumentando en los años venideros. Esto influirá negativamente tanto en las temperaturas como en la calidad del aire y por ende en la salud de la ciudadanía. Es por esto por lo que un cambio es realmente necesario para evitar que el nivel de calidad de vida de la ciudad se vea reducido.

Son por todas nuestras conclusiones por las que abogamos por un diseño verde de la ciudad, no sólo para los nuevos barrios que se creen, sino que se debe aplicar en toda la ciudad un diseño verde, Griffiths y Sovacool (2020).

Desde nuestro punto de vista este cambio tiene que venir dado por estos principales factores:

- El punto más obvio es reducir las emisiones de gases contaminantes como el Dióxido de Nitrógeno, el cual hemos analizado. Para ello hay que seguir apostando por medidas como Madrid central, que limita el acceso de coches contaminantes al centro de la ciudad, una de las zonas con mayor efecto ICU. Del mismo modo hay que incentivar el uso de transporte público, así como el de vehículos de bajas emisiones. Es decir, priorizar el transporte público y desincentivar el uso de vehículos privados y promover el acceso a facilidades (supermercados, ocio, cultura...) por proximidad

- Seguir ampliando las zonas verdes y azules de la ciudad, no sólo en los exteriores y en barrios de nueva construcción, sino que son tremendamente necesarios en cada uno de los núcleos urbanos. Apoyando el reverdecimiento urbano y la jardinería comunitaria.

- En zonas de nueva construcción apostar por zonas abiertas como avenidas y edificios de no mucha altura, limitaría el efecto cañón. Así como invertir en diseños de alta eficiencia energética en los edificios como el uso de materiales que no absorban e irradien calor, para paliar los efectos negativos y no incrementarlos, sobre todo en épocas invernales y estivales.

- Aunque estemos en un punto en que se hacen campañas a favor de medidas como las propuestas hay que seguir realizando una concienciación pública sobre el medio ambiente local a través de la educación y la divulgación en todos los estratos de la sociedad.

La implantación de medidas de este estilo a nivel gubernamental sería beneficiosa tanto para la ciudadanía como para la economía del propio Ayuntamiento, ya que la eficiencia energética es cada vez más y más rentable a largo plazo, Frérot (2014).

7.3. Trabajo futuro

Como ya hemos citado antes, previsiblemente el efecto ICU se vaya incrementando en los próximos años. Es por esto por lo que creemos que un análisis de los años previos sería de gran utilidad para poder prever como se comportará en el futuro.

Este objetivo es relativamente sencillo ya que en el propio Portal de Datos Abiertos del Ayuntamiento se muestran accesibles ficheros de años pasados.

Del mismo modo sería una gran idea ver como afectó el virus SARS-CoV-2 a la ciudad. Debido al largo periodo de cuarentena, en el que casi no había contaminación por parte de vehículos y la calidad del aire fue mucho mejor. Verlo comparado con el 2021 podría esclarecer en cómo se comporta la ciudad en escenarios de 0 emisiones o cercanos a estas.

Otro trabajo curioso que se podría plantear a raíz de este sería una mayor precisión a la hora de contar las zonas verdes. Puesto que nuestro trabajo asignaba la masa arbórea del distrito entero a las estaciones, no llega a ser tan realista como si se cogieran datos por proximidad a estas. Esto podría arrojar resultados más precisos que permitirían explicar el efecto ICU con mayor claridad.

En cuanto a la parte de visualización, si se consiguiese un Dataset más amplio el cual no tuviese tantos valores nulos, con las herramientas de visualización de R y RStudio se podrían conseguir mapas más elaborados. Hay bibliotecas que permiten realizar plots interactivos, como poder representar los datos anuales como GIF´s, imágenes animadas o que el propio usuario pudiese interactuar con un entorno para obtener distintas representaciones.

Estas son muchas de las opciones que ofrece este trabajo para futuras mejoras, que con ayuda de los conjuntos de datos del Ayuntamiento pueden llegar a explicar muchos fenómenos de la ciudad de Madrid.

Introduction

The average temperature in large cities is rising every year, setting new records. However, in small cities and towns not so far away from these areas, this phenomenon does not happen. Why is this the case? The answer lies in the *Urban Heat Island*(ICU onwards).

This problem is a consequence of a variety of factors. Starting with the large number of buildings and building materials, which absorb heat and radiate it slowly. In addition to the pollution generated by vehicles and air conditioning systems. Also, the scarcity of green and fluvial areas increases this phenomenon. This, added to climate change, increases temperatures, which implies an increase in the population's energy consumption. But it not only affects economic issues, but also has an impact on citizens' health, aggravating pathologies, worsening mental health, etc. Sierra (2021).

Madrid, with 3,305,408 people (INE 2021), is the most populated city in Spain and the second most populated city in the European Union[1]. In addition, the M-30, A-2 and M-40 are high traffic roads. All this places it among the cities with the worst air quality in Spain[2], which means that the ICU effect is intensified.

Thanks to the Open Data Portal of the Madrid City Council[3] we can access to a large dataset that stores information about the municipality. For this TFG we will use data from the different meteorological and air quality stations of the districts that compose the city. After processing the data, we will analyze how the ICU effect is distributed in Madrid using Machine Learning (ML onwards) algorithms, such as KMeans, and representations in graphs such as dendrograms, and more visual forms such as maps.

7.4. Motivation

The main idea that led us to carry out this work was the huge difference in temperatures between the center of Madrid and the suburbs and the desire to explain this event, since in a matter of kilometers the temperature changes on average about 3^{o}C. But the most striking thing was that on days of extreme temperatures, as can

[1]Link Population Data European Union
[2]Article Cities with the worst air quality
[3]Link to the Open Data Portal of the Madrid City Council

happen in winter, the difference reaches more than 15ºC, as can be seen in Figure 7.1.

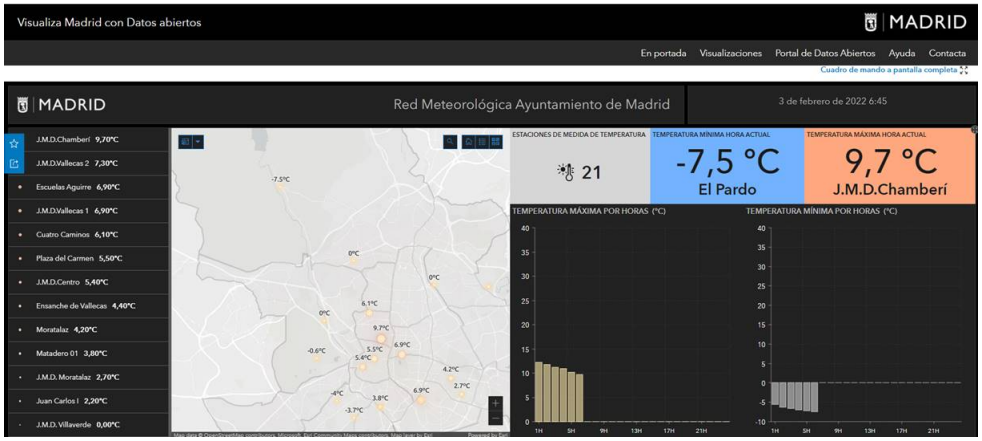


Figure 7.1: Difference of 17.2ºC between Madrid center and El Pardo.

In addition, the fact that this is the city in which we live and that this event called ICU directly affects us and our quality of life, prompted us to analyze the reasons for these differences.

Apart from the above, we also have a personal interest. Throughout our university education at the faculty, we have developed an interest in Big Data and ML, which encouraged us to do this work.

7.5. Objectives

The main objective of the work is the application of ML algorithms to detect which areas behave in a similar way in terms of meteorological and urban variables in Madrid and the reasons that produce it. The objectives could be summarized as follows:

- Perform cleaning, formatting and filtering techniques of large-scale data, collected from the Open Data Portal of the Madrid City Council, in order to obtain homogeneous data that can be analyzed.
- Apply algorithms from the world of Data Science to be able to perform different classifications of weather stations.
- Use of data visualization tools to illustrate these groupings for easier interpretation of the results.
- Draw conclusions from classifications through visualization and propose solutions.

Data collection, data exploration and processing, choice of algorithms, visualization and analysis are fundamental in Data Science.

Another objective, although personal, is to learn to program in a new language for both of us, such as R, as well as the necessary tools for cleaning, classification and

visualization through the use of libraries in RStudio and to understand the context of temperature differences in our city of residence, Madrid.

7.6. Work plan

Thanks to the knowledge we have acquired in the subjects taught and, in the projects, carried out during the different Degrees, especially in Software Engineering, we have been able to make a planning that we have considered quite useful and decisive. Being a small team, two students and a tutor, and we have not been able to apply the agile development methods as we have been taught, we have tried to adapt it to the size of the team.

Our work has been divided into two main branches, a memory draft and the code.

- **Memory draft**: Since we have been warned from the beginning of the importance of the final report, we saw it convenient to generate a document in which we could write all the information we collected, explanations of the data formats and their treatment, explanations of the code and observations and links of interest. This document has been very useful when writing the report, as it has facilitated access to the information we have collected, as well as the steps we have followed to get to the end of the work.

- **Code**: As with all software development, it is always necessary to use a repository to control the versions that are released throughout the process. So much so that we decided to create one on the **GitHub**[4] platform where we have been uploading the code belonging to the different objectives set.

7.6.1. Development language: R

The approach we wanted to give to the work was more statistical with parts of analysis and visualization on large data sets so in the research phase we contemplated different languages in which to program, but as we wanted to emphasize a statistical vision, we decided to use R. Furthermore, since the application of ML algorithms was going to be very relevant for the classification of weather stations, we reaffirmed our decision. This is because there are a large number of libraries, and even in R base, where functions implementing such algorithms are included.

7.6.2. Development environment: RStudio

Once the language is chosen, choosing a development environment is much easier. There were many options, from *VisualStudio*, *Geany*, *RKWard*, etc. Apart from the fact that most of the online tutorials used RStudio, we decided to be guided by the recommendation of our tutor, Sonia, so this work has been completely developed with the *IDE RStudio*. This has been very useful since it shows us all the objects in

[4]Enlace al repositorio, https://github.com/Gmene00/TFG-Madrid-Isla-de-calor

the *workspace* as well as the commands executed either in *scripts* or in the console it provides, as well as a viewer with the available and installed packages. This tool has a graphics viewer, which, together with the extensive number of available libraries, makes the visualization intuitive, simple and fast.

These are the main libraries used in the work:

- Package *Tidyverse*, is a collection of packages and libraries designed for the *Data Science*, which allows us to manipulate the data for a better interpretation both in tables and graphs. From this collection we have used the following packages:
  - Library *ggplot2*: allows to create graphs declaratively. It is enough to provide it with the data and the aesthetic variables and it easily generates the figure. As a complement, there is *ggpubr* that increases the possibilities to show the data in a more elegant way.
  - Library *dplyr*: allows you to manipulate data in a simple way, especially thanks to the function *select()*, which is able to return subsets of larger ones in a single statement.
  - Library *tidyr*: when cleaning data, many of its functions that eliminate null values are useful, as well as when formatting, it offers functions to split cells, reshape data frames, etc.
- Library *cluster*: As defined by the library itself, it provides methods for the analysis of *clusters*.
- Extra fact library: contains functions that apply ML algorithms and facilitate visualization, such as the elbow diagram.
- Library *osmdata*: allows to download and import information from OpenStreetMap as sp or sf objects. It allows us to make the maps.
- Library *sp*: provides classes and methods to deal with spatial data.
- Library *ggmap*: adds an additional cartographic layer to the already known graphics, those extracted with own elaboration.

7.6.3. Development method: Waterfall

To continue with a correct development of the *TFG*, we see convenient to specify how we have been structuring, planning and controlling the project over time. To do this we have tried to follow the steps of agile development methodologies, although these are recommended for groups of between five and nine people, we adapt it to our needs to be able to structure ourselves in an appropriate way that we explain below.

The objectives of the work are dependent on each other, i.e., we cannot apply the *ML* algorithms without first having cleaned and formatted the data, just as we cannot visualize the results in the form of graphs or maps without first having applied

the algorithms themselves. This made us choose the *Waterfall* or *Cascade* methodology, which is characterized by dividing the project into differentiated functions, which for us have been the objectives.

This methodology requires that before moving on to the next function or objective, the project must be reviewed and in case of finding faults it must be corrected. This is where the tutor of the *TFG*, Sonia, comes in, with whom we have arranged meetings approximately every two weeks, mostly face-to-face. Where we were oriented in each stage of the *TFG* and she checked that the set objectives were being achieved.

Another advantage of following this model is that, since the objectives were well defined very early in the development process, planning has been simple.

In addition, for greater organization and control over dates, as well as task distribution, we have used a Gantt diagram that has allowed us to be more aware of the work done up to each meeting, as well as the pending objectives. As a summary of the planning, Figure 7.2 shows what the process has been like.

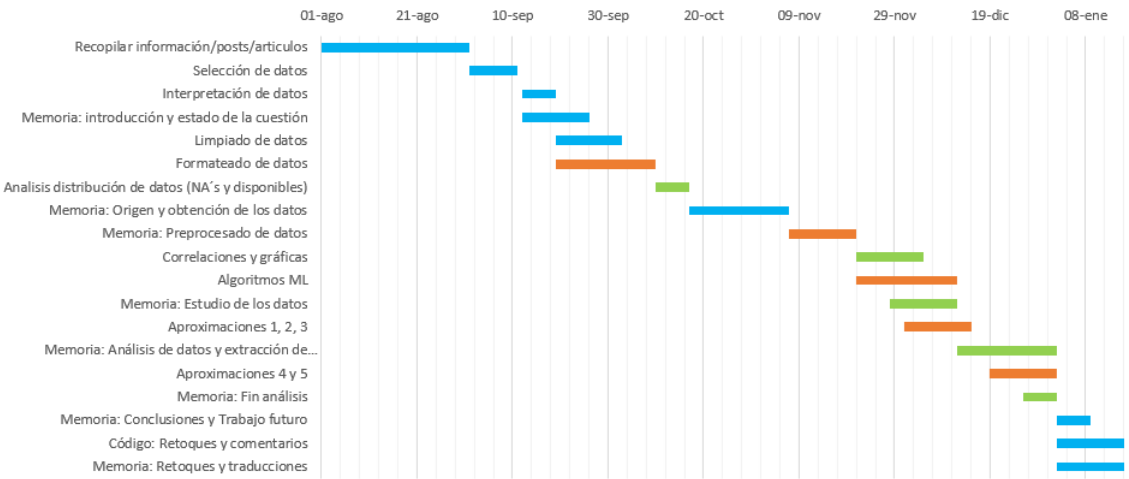


Figure 7.2: Work; blue-*Both*, green-*José*, orange-*Gonzalo*

7.7. Memory content

This memory consists of the following sections:

1. Introduction: this chapter presents the problems, objectives and mechanisms to achieve them.

2. State of the art: during this chapter the different areas of Madrid and its heat signature will be presented, contextualizing and classifying each of these areas, apart from the current status of various works that address their study. The main differences of this project with other related projects will also be presented.

3. Origin and collection of data: this section will present the source from which the data were collected and the methodology used to obtain them. where

the data have been extracted, and the methodology used to obtain them. In addition, the description of the data in its original form is also explained.

4. Preprocessing of the data: where the methods used for data processing and cleaning will be explained in order to methods used for data treatment and cleaning, in order to achieve a suitable format for them adequate format for its study, and thus facilitate the extraction of information. information. The result obtained after the transformation is also shown.

5. Study of the data: the different techniques and methods for the treatment of null values and outliers are shown. treatment of null values and unexpected values or outliers, with the objective of not losing too much not to lose too much relevant information and to eliminate information that may introduce noise in the results introduce noise in the results.

6. Data analysis and information extraction: in this section we present the algorithms that have been used for the creation and selection of models and the techniques that can be used to improve their predictions. Also, the comparison between the different models will be presented.

7. Conclusions and future work: this section lists the conclusions reached after the results obtained during the previous section. the results obtained during the previous section are listed in this section. In addition, possible future possible future work that could be carried out as a result of this project are also described.

Conclusions and Future Work

7.8. Conclusions

In view of the results obtained during the course of this work and taking into account the objectives defined at the beginning, we have reached the following conclusions:

O1. The data had to be cleaned, formatted and filtered to be collected in a single file. It was necessary to make a change in the format that is included in point 4, which increased the difficulty at the time of doing it. Removing this difficulty, it has been possible to unify it in a single file thanks to the R tool and libraries.

From the fulfillment of this objective, we can conclude that the format provided by the City Council itself could be improved. It establishes months with more days than they correspond to, such as February with 31, which can lead to confusion. More intuitive and manageable ones can be proposed, such as the one we have done in the cleaning, which facilitates the compression of these.

O2. It has been possible to apply algorithms and methods accepted in the world of Data Science in order to be able to different classifications of the stations. From these we have obtained new conclusions:

- Green areas have a significant influence on temperatures. Large city parks with abundant trees and blue areas such as ponds, fountains or the river itself are grouped in the same clusters. Temperatures are softened in these parts of the city, reducing the ICU effect.

- Urban centers are included in another group. They are characterized by high temperatures above those mentioned in the previous point, as well as being quite polluted areas. In these groupings we realize that the canyon effect, mentioned in the State of the Question, is a proven fact. Areas where buildings abound and wind speeds are lower, accumulate heat in the environment. This directly influences temperatures and therefore the ICU effect.

- The areas near the M-40 and M-30 highways consolidate another set. Mainly due to their construction materials, which absorb and radiate heat more slowly. As in the previous point, these areas are quite polluted, but, in general, as they are open, they limit the canyon effect.

- Another of the main conclusions is the urban overcrowding in Madrid. Only the outermost districts are those with large masses of trees that soften the temperatures. The closer we get to the center of the city, the greener areas become noticeably smaller, with exceptions such as Retiro Park or Parque del Oeste.

- The presence of buildings, roads and other urban elements, in addition to limiting the flow of air needed to reduce temperatures, are composed of materials that increase them. These absorb heat and radiate slowly, making temperature differences more noticeable between urban and outdoor areas with fewer buildings and green areas.

- Although the ICU effect is notorious throughout the year, it is intensified during the winter and summer seasons. In other words, temperatures are more extreme in summer, which, together with the use of electrical appliances to alleviate the high temperatures, increases the ICU effect. And during the winter, temperatures are unseasonal, with differences of around 7-8^{o}C on average between the center and the outdoor green areas.

- Taking into account the above conclusions, we can locate the ICU effect mainly in central Madrid, where there is a lack of green areas, a greater canyon effect and higher pollution levels than in the rest of the city. The urban heat island is dissipating towards the north, south and east, which is where neighborhoods built in the late twentieth century and early twenty-first century begin to appear. Its main characteristics are open spaces and large avenues (limiting the canyon effect), as well as the creation of large green areas such as the Almudena Cemetery, the Cuña Verde, Juan Carlos I Park and Felipe VI Park, among others.

O3. It has been possible to make use of data visualization tools to illustrate these groupings for an easier interpretation of the results, using graphs and maps. for an easier interpretation of the results, by means of graphs and maps, which are important to confirm and observe the Heat Island Effect.
We have concluded that roads and pollutants have a negative effect on temperatures and air quality, of which we can establish a direct relationship. Similarly, it is appreciated with the outside of the city and green areas and better quality.

7.9. Recommendations

In view of the conclusions obtained and the increasing climate change, the ICU effect is expected to continue to increase in the coming years. This will have a negative influence on both temperatures and air quality, and therefore on the health of citizens. This is why a change is really necessary to prevent the city's quality of life from being reduced.

It is because of all our conclusions that we advocate a green design of the city, not only for the new neighborhoods to be created, but a green design, Griffiths y Sovacool (2020), should be applied throughout the city.

From our point of view, this change must be brought about by these main factors:

- The most obvious point is to reduce emissions of polluting gases such as nitrogen dioxide, which we have analyzed. To this end, we must continue to support measures such as central Madrid, which limits the access of polluting cars to the city center, one of the areas with the greatest ICU effect. Similarly, we must encourage the use of public transport as well as low-emission vehicles. In other words, prioritize public transport and discourage the use of private vehicles and promote access to facilities (supermarkets, leisure, culture...) by proximity.
- Continuing to expand the city's green and blue zones, not only in the outdoors and in newly built neighborhoods, but are tremendously needed in every urban core. Supporting urban greening and community gardening.
- In new construction areas, open areas such as avenues and low-rise buildings would limit the canyon effect. As well as investing in energy-efficient designs in buildings such as the use of materials that do not absorb and radiate heat, to mitigate the negative effects and not increase them, especially in winter and summer seasons.
- Although we are at a point where we are campaigning for measures such as those proposed, we must continue to raise public awareness of the local environment through education and outreach to all strata of society.

The implementation of measures of this style at the government level would be beneficial both for the citizens and for the economy of the City Council itself, since energy efficiency is becoming more and more profitable in the long term, Frérot (2014).

7.10. Future Work

As mentioned above, the ICU effect is expected to increase in the coming years. This is why we believe that an analysis of previous years would be very useful to be able to foresee how it will behave in the future. This objective is relatively simple, since the City Council's own Open Data Portal provides access to files from previous years.

It would also be a great idea to see how the SARS-CoV-2 virus affected the city. Due to the long quarantine period, where there was almost no pollution from vehicles and the air quality was much better. Seeing it compared to 2021 could shed light on how the city behaves in zero-emission or near-zero-emission scenarios.

Another curious piece of work that could follow on from this would be a more accurate count of green space. Since our work assigned the tree mass of the entire district to the stations, it is not as realistic as if data were collected by proximity to the stations. This could yield more accurate results that would allow us to explain

the ICU effect more clearly.

As for the visualization part, if a larger Dataset were obtained which did not have so many null values, with the visualization tools of R and RStudio more elaborated maps could be obtained. There are libraries that allow interactive plots, such as being able to represent the annual data as GIFs, animated images or that the user himself could interact with an environment to obtain different representations.

These are many of the options offered by this work for future improvements, which with the help of the City Council's datasets can come to explain many phenomena in the city of Madrid.

Contribuciones Personales

En el capítulo de 1 en el punto 1.3.3. hacemos mención al Método de desarrollo en Cascada y cómo se han distribuido las tareas a lo largo del trabajo, Figura 7.3

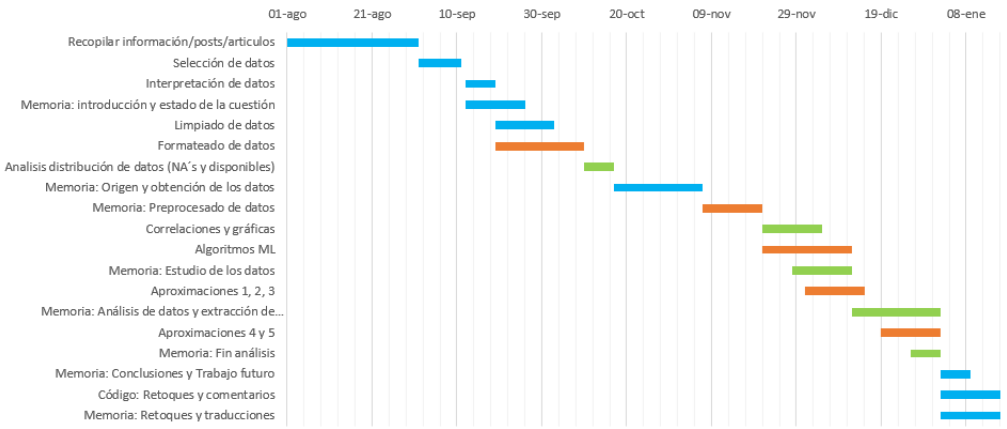


Figura 7.3: Trabajo; azul-*Ambos*, verde-*José*, naranja-*Gonzalo*

A continuación, detallaremos los trabajos realizados por cada uno de nosotros.

Gonzalo Meneses Vicente

Desde mi punto de vista la carga de trabajo de ambos ha estado equilibrada, aunque eso sí, creo que yo me he centrado más en la parte del código.

Como se puede ver en el diagrama la primera parte del trabajo ha sido conjunto. Esta se ha conformado de investigación, recopilación de información, papers, artículos, etc. Cada uno estuvo informándose de un tema del trabajo. En mi caso fue la parte de meteorología.

Estuve investigando cómo la meteorología y la disposición urbana (zonas verdes, carreteras, etc.) afectaban a las temperaturas de una ciudad y por lo tanto de como influían en una Isla de Calor Urbana. En este punto leí varios informes al respecto como fueron:

- El efecto cañón (explicado en el capítulo 2), que expone como la velocidad y dirección del viento se ve afectada por los distintos elementos urbanos (edificios, carreteras, etc.) aumentando las temperaturas en dichas zonas, Lipp (2014).

- El efecto de las zonas verdes y azules en las temperaturas, así como los grandes parques y zonas forestales, con los informes Arellano Ramos y Roca Cladera (2018) y Salinas 2020.
- Informes sobre carreteras, su absorción de calor y posterior irradiación debido a sus materiales Xu et al. (2021), y el efecto de la contaminación por parte de estas, Piracha y Chaudhary (2022).
- El efecto de la humedad y precipitaciones en las temperaturas de una ciudad, Tejedor et al. (2016).
- Y por último también investigué posibles soluciones para resolver o al menos mitigar la problemática de las temperaturas extremas causadas por el efecto ICU, Griffiths y Sovacool (2020).

Una vez puesto en contexto como íbamos a orientar el trabajo, nos expusimos el uno al otro toda la información recogida para llegar a un punto común.

Después estuvimos buscando entre los dos los conjuntos de datos a tratar, llegando al Portal de Datos Abiertos del Ayuntamiento.

Una vez obtenidos los datos con los que trabajar, mis tareas con el código fueron:

- Creación del repositorio en GitHub y establecimiento del formato de directorios.
- Limpiado de los datos de meteorología explicado en el capítulo 4. Esto incluye el script del GitHub limpieza.R en el que también se limpian los datos de contaminación, pero por parte de José.
- Una vez limpiados los datos nos encargamos de pensar y diseñar el nuevo formato, con el fin de que fuese más legible para el posterior análisis. De nuevo yo me encargué de la parte de meteorología, aunque esta vez tuve que ayudar a José a establecer el formato para los de contaminación.
- En cuanto a las gráficas, hice la base del fichero CodigoCorrelaciones _GraficasLineales.R sacando unas primeras visualizaciones muy sencillas.
- Los algoritmos de Machine Learning en un principio los estuvimos investigando los dos, pero a la hora de aplicarlos sobre las 5 aproximaciones las codifiqué yo, en el archivo clustering.R. Eso sí la planificación de estas fue, otra vez, parte de los dos.
- La parte de visualización empezó haciéndola José, con una base sencilla, pero a la hora de representar las visualizaciones el código fue realizado por mí mediante MapaMadrid.R.

Y pasando a la parte de la memoria, aquí he de decir que la parte más densa corresponde al trabajo de José. Mis aportaciones han sido:

- Trabajo conjunto con José de la Introducción. Aunque el diagrama de Gantt fui yo el que lo iba actualizando con cada una de las tareas completadas.

- En la parte de Estado de la cuestión los puntos 2.1.1., 2.2., 2.3. al completo, 2.4. y 2.5. fueron redactados por mí.

- En el capítulo 3 participamos ambos, ya que esta parte tuvimos que decidirla entre los dos.

- La redacción del punto 4 también fue realizada por mí, a excepción del 4.2.2. A este punto hay que sumarle que el diagrama de flujo de la limpieza, que también es de mi autoría.

- El capítulo 7 fue redactado al completo por mí, tanto las conclusiones, recomendaciones y trabajo futuro.

- Por último me encargué de la traducción de los capítulos 7.3 y 7.7.

Como resumen de mi trabajo puedo decir que me he centrado más en la parte del código, aunque José también ha trabajado en él. Aun así, también he formado parte del análisis ya que para redactar las conclusiones (capítulo 7) tuve que examinar todas las gráficas obtenidas por José y sus interpretaciones. Del mismo modo, al generar los mapas de Madrid y aplicar los algoritmos de ML he participado de forma activa en el proceso de Estudio de los datos (capítulo 5) y Análisis y extracción de información (capítulo 6).

Como valoración del trabajo de José, ha participado activamente en el código en la parte de limpieza de los datos de contaminación, primeras visualizaciones de los mapas y la gran mayoría de las gráficas.
Y sobre todo se ha encargado de la memoria. Tanto de la redacción de los puntos 5 y 6, que considero más densos y complicados de redactar y plasmar, como correcciones de todos los puntos (incluidos los míos), organización de la memoria, tablas, imágenes, etc.

José Francisco García Ruiz

Durante el desarrollo del TFG la carga de trabajo ha sido equilibrada, valorando en un primer momento los puntos fuertes de cada uno de los componentes decidimos que por experiencias personales y laborales la carga de gestión y organización fuera llevada por mí, mientras que la carga de código fuese llevada por Gonzalo.

Partiendo al igual que mi compañero del diagrama Figura7.3, mi parte de investigación, recopilación de información, papers, artículos, etc. fue sobre la parte de contaminación.

Tras hacer un análisis de las necesidades para el trabajo tanto de investigación como de capacidades a desarrollar y lenguajes y entornos de trabajo que íbamos a realizar, empezamos a investigar cada uno sobre el tema a tratar, poniendo en común todos estos datos en la siguiente reunión que tuvimos. De esta reunión anterior sacamos que el entorno donde íbamos a trabajar iba a ser RStudio, con R como lenguaje y que la base de datos de la que nos íbamos a nutrir sería el portal de datos abierto del Ayuntamiento de Madrid.

Durante el desarrollo de todo el proyecto fuimos gestionando reuniones periódicas con la tutora para ir planteándole propuestas y mostrarle el desarrollo del trabajo. Se puede decir que el trabajo tuvo una primera parte de informarse y aprender sobre el código y metodologías que íbamos a tratar sacando referencias de librerías, después tuvimos una parte donde nos tocó desarrollar código.

En esta parte Gonzalo fue quien dirigió los pasos para el desarrollo total del código. Ayudándome siempre que lo necesitaba en el desarrollo del código. También se gestionó las diferentes plataformas como *Drive* o *GitHub*.

Aquí el procedimiento que solíamos hacer era muy parecido, cuando tocada desarrollar alguna parte del código solíamos dividirnos la tarea. Por poner un ejemplo: Gonzalo se encargaba de desarrollar limpieza.R de los datos de meteorología mientras yo empezaba a investigar sobre cómo realizar la representación de mapas que usamos en el trabajo. Cuando Gonzalo terminaba nos intercambiamos la información y yo adaptaba el código a los datos de contaminación y Gonzalo terminaba el desarrollo de los mapas.

Aquí hay un punto de inflexión durante el trabajo, por necesidades del análisis tuvimos que cambiar el formato de los datos, algo con mucha carga en lo que Gonzalo consigue encontrar una solución, planteándola y pudiéndola desarrollar el para los datos de meteorología y yo con su ayuda para los datos de contaminación. Tras esto estuvimos informándonos sobre los métodos K-Means y Jerárquico.

Donde estuvimos investigando los dos y Gonzalo se encargó de desarrollar la codificación del código.

Por último, quedaba la representación cosa que me encargaba mientras Gonzalo realizaba el punto anterior. Aquí me encontré con varios problemas, llegando a realizar representaciones sencillas de los distritos de Madrid y representado las estaciones en él. Seguí investigando con *Google Maps* y fuentes abiertas, pero no llegando a mostrar más que vectores dibujados, por lo que Gonzalo me ayudo consiguiendo el mostrar las figuras finales como esta del trabajo. Figura6.10

Como explicaba anteriormente cada uno se hace cargo de un conjunto de datos, contaminación en mi caso de donde saco información de los siguientes puntos. Arellano Ramos y Roca Cladera (2018) de donde nos servimos tanto Gonzalo como yo para informarnos sobre todo el efecto de isla de calor. Ballester (2005) donde nos informamos sobre el impacto de los contaminantes sobre la salud humana y el impacto en el ecosistema, de aquí también extraemos los principales gases contaminantes para el estudio a realizar.

Y otros artículos como Lipp (2014), Piracha y Chaudhary (2022) donde nos informamos y sacamos conclusiones sobre el tema a tratar. Tras el análisis inicial también nos servimos de las siguientes referencias para poner en común la información sacada por mi compañero y por mí, Collado Mamblona (2022), Trabajo fin de grado donde estudia la relación entre meteorología y contaminación en la ciudad de Madrid.

Por último, de la memoria me encargue de pasar toda la información que teníamos hasta entonces y darle una estructura para empezar a trabajar en *Overleaf*. También fui el encargado de repasar y verificar que el formato estuviera bien y correcciones que tuvimos con la tutora. Y me encargue de subir todas las imágenes necesarias para el trabajo, aparte de realizar todas las tablas de este. Aparte

desarrolle los siguientes puntos:

- Capitulo introducción 1 Esto fue un trabajo conjunto que fuimos realizando desde el inicio del trabajo, cosa que nos ayudó mucho a la hora de empezar a redactarlo en *Overleaf*.
- Capitulo estado de la cuestión 2, toda la parte de contaminación fue redactada por mí.
- Capitulo origen y obtención de los datos 3 esta parte es común.
- Capitulo 5 y 6 fue realizado por mí, pero Gonzalo lo reviso añadiendo cosas que él había analizado o comprobado durante el trabajo.
- Por último me encargué de comprobar todas las figuras, tablas y su correcta definición.

Como resumen de mi trabajo me he encargado más de la parte de trascribir lo que íbamos desarrollando en la memoria, la búsqueda de información y la comunicación con la tutora, aportando y desarrollando también parte del código, tuve que examinar el código desarrollado por Gonzalo y adaptarlo a los datos de contaminación, realizando también todas las gráficas. A la hora de hacer la memoria he sido el encargado de ver cómo funcionaba *Overleaf* y transmitírselo a mi compañero, encargándome de dejar todas las figuras y tablas en su sitio óptimo, no siendo posible todos los casos por el funcionamiento de *Overleaf*.

Como valoración del trabajo de Gonzalo, ha sido una persona resolutiva en el desarrollo del software, ayudándome las veces que así lo he necesitado, cosa a valorar y agradecer. También se ha encargado de la memoria, desarrollando los puntos que a él le tocaban. Valorando que ha tenido un desarrollo de este trabajo muy bueno y completo. Informándose en todo momento de lo que realizaba para saber también sobre los puntos que él no tenía que realizar.

Bibliografía

Arellano Ramos, B. y Roca Cladera, J. Áreas verdes e isla de calor urbana. En *Libro de proceedings, CTV 2018: XII Congreso Internacional Ciudad y Territorio Virtual:"Ciudades y Territorios Inteligentes": UNCuyo, Mendoza, 5-7 septiembre 2018*, páginas 417–432. Centre de Politica de Sol i Valoracions, CPSV/Universitat Politècnica de ..., 2018.

Ballester, F. Contaminación atmosférica, cambio climático y salud. *Revista Española de salud pública*, vol. 79, páginas 159–175, 2005.

Collado Mamblona, A. Relación entre meteorología y contaminación en la ciudad de madrid. 2022.

Frérot, A. Economía circular y eficacia en el uso de los recursos: un motor de crecimiento económico para europa. *Cuestión de Europa*, vol. 331, páginas 1–10, 2014.

Griffiths, S. y Sovacool, B. K. Rethinking the future low-carbon city: Carbon neutrality, green design, and sustainability tensions in the making of masdar city. *Energy Research & Social Science*, vol. 62, página 101368, 2020.

Lipp, D. El cañón urbano su incidencia en la contaminación del aire. En *Actas Congreso Internacional de Geografía*, vol. 75, páginas 123–128. 2014.

Piracha, A. y Chaudhary, M. T. Urban air pollution, urban heat island and human health: A review of the literature. *Sustainability*, vol. 14(15), página 9234, 2022.

Salinas 2020, J. A. G. Criterios para la planificación y diseño de los corredores fluviales urbanos para la mitigación de la isla de calor. ????

Sierra, R. M. Características espaciales del impacto de la isla de calor y su relación con el consumo eléctrico en hermosillo, sonora. En *3er Congreso Internacional de Arquitectura y Diseño (CIAD 2021)*. 2021.

Tejedor, E., Cuadrat, J. M., Saz Sánchez, M. Á., Serrano Notivoli, R., López, N. y Aladrén, M. Isla de calor y confort térmico en zaragoza durante la ola de calor de julio de 2015. 2016.

Xu, L., Wang, J., Xiao, F., Sherif, E.-B. y Awed, A. Potential strategies to mitigate the heat island impacts of highway pavement on megacities with considerations of energy uses. *Applied Energy*, vol. 281, página 116077, 2021.