

Predicción de la contaminación ambiental mediante ondas de radio frecuencia

Guillermo Calvo Arenaza

MÁSTER DE MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS. FACULTAD
DE ESTUDIOS ESTADÍSTICOS
UNIVERSIDAD COMPLUTENSE DE MADRID



Trabajo Fin Máster

2022-2023

Director/es y/o colaborador/es:

Director: Daniel Gómez González
Colaborador: José Luis Vázquez Poletti

Resumen

En este trabajo se estudia la posibilidad de poder predecir la contaminación ambiental de partículas en suspensión, así como la existencia de ciertos gases de una zona mediante la presencia de ondas de radio frecuencias asociadas a actividades telefónicas, sensores y aviación.

De esta forma extrayendo los mensajes transportados en dichas frecuencias se ha podido realizar un conteo de las ruedas de vehículos, al igual que los teléfonos y aviones cercanos con los que intentar predecir la contaminación ambiental.

Palabras clave

Contaminación ambiental, ondas de radio frecuencia, telefonía, aviación, sensores, data mining, series temporales.

Abstract

This project studies the possibility of being able to predict the environmental pollution of suspended particles and the presence of certain gases in an area through the presence of radio frequency waves associated with telephone activities, sensors and aviation.

By extracting the messages carried on these frequencies, it is possible to count the wheels of cars, nearby telephones and nearby airplanes in order to try to predict environmental pollution.

Keywords

Environmental pollution, radio frequency waves, telephony, aviation, sensors, data mining, time series.

Índice general

Índice	I
Agradecimientos	v
Dedicatoria	1
1. Introducción	1
1.1. División institucional de las franjas de frecuencias electromagnéticas	1
1.1.1. 433 MHz	4
1.1.2. 1090 MHz	6
1.1.3. 930 MHz	7
1.2. IMSI catchers	10
1.3. Estación meteorológica	11
1.3.1. PMS7003	12
1.3.2. MQ 135	13
2. Objetivos	15
3. Metodología	17
3.1. Recopilación de datos	17
3.1.1. Ejemplos datos IMSIS	18
3.1.2. Ejemplos datos ambientales	18
3.1.3. Dificultades en la recopilación de datos	19
3.2. Bases de datos	19
3.3. Agregación de los datos	19
3.4. Limpieza y división de la base de datos	21
3.5. Metodología SEMMA	21
3.6. Selección de modelo	21
3.7. Metodología de selección de modelo	23
3.8. Software empleado en los modelos	23
4. Desarrollo	25
4.1. Descriptiva de los datos	25
4.2. Estacionalidad de las series temporales	28
4.3. Causalidad de Grangers de las series temporales	28
4.4. Ciclos de las series temporales	29
4.5. Modelado de series temporales ARIMA	31
4.5.1. Modelado de ARIMA	33

4.6.	Modelado de series temporales multivariante	34
4.6.1.	Modelado de VAR	35
4.7.	Modelado con redes neuronales LSTM	37
4.8.	Comparación modelos	40
4.9.	Ensamble de predicciones	41
4.10.	Selección del mejor modelo	42
5.	Conclusiones	45
	Bibliografia	50
	A. Bases legales del trabajo	51

Agradecimientos

No podría empezar a escribir esta memoria, sin agradecer en primer lugar a la Oficina de Software Libre UCM⁸ por su magnífica labor divulgativa así como la realización de los talleres que han inspirado la idea de este TFM y la ayuda prestada para llevar a cabo este trabajo.

En segundo lugar, agradecer también al departamento de redes de Orange España por la ayuda prestada a la hora de comprender el funcionamiento de las antenas objetivo de este trabajo.

Este trabajo no habría sido posible sin la ayuda de:

- José Luis Vázquez Poletti (Oficina de Software Libre)
- Jesús Alarcón Roldan (Orange España)
- Jorge Torres Fernández (Orange España)
- Jaime Sergio Martínez Peligros (Estudiante de la Facultad de Informática UCM)
- Grupo de RadioHacking FDI (Facultad de Informática UCM)⁷
- Gonzalo José Carracedo Carballal (Doctorando en Astrofísica UCM)⁴

Dedicatoria

44 65 64 69 63 61 64 6F 20 61 20 6D 69 20 6D 61 64 72 65 20 4D 61 72 74 61 2C 20 70
6F 72 20 73 65 72 20 75 6E 20 65 6A 65 6D 70 6C 6F 20 64 65 20 73 75 70 65 72 61 63 69
6F 6E 20 64 69 67 6E 6F 20 64 65 20 73 65 67 75 69 72 2E 20 0A 0A 44 61 20 69 67 75 61
6C 20 63 75 61 6E 74 6F 73 20 62 61 63 68 65 73 20 74 65 20 70 6F 6E 67 61 20 6C 61 20
76 69 64 61 2C 20 73 69 65 6D 70 72 65 20 6C 6F 73 20 73 75 70 65 72 61 20 63 6F 6E 20
75 6E 61 20 62 6F 6E 69 74 61 20 79 20 61 6C 65 67 72 65 20 73 6F 6E 72 69 73 61 2E 20
0A 0A 4E 6F 20 65 78 61 67 65 72 6F 20 63 75 61 6E 64 6F 20 61 66 69 72 6D 6F 20 71
75 65 20 4E 4F 20 65 78 69 73 74 65 6E 20 70 61 6C 61 62 72 61 73 20 6E 69 20 69 64 69
6F 6D 61 73 20 63 61 70 61 63 65 73 20 64 65 20 64 65 73 63 72 69 62 69 72 20 61 20 6C
61 20 6D 65 6A 6F 72 20 61 6D 61 63 68 75 20 64 65 6C 20 70 65 71 75 65 F1 6F 20 62 61
72 72 69 6F 20 62 69 6C 62 61 69 6E 6F 20 70 65 6E 69 6E 73 75 6C 61 72 20 28 63 6F 6D
75 6E 6D 65 6E 74 65 20 63 6F 6E 6F 63 69 64 6F 20 63 6F 6D 6F 20 72 65 69 6E 6F 20
64 65 20 45 73 70 61 F1 61 20 3A 50 29 20 0A 0A 53 65 72 65 20 75 6E 20 68 69 6A 6F 20
74 65 72 72 69 62 6C 65 20 61 20 6C 61 20 68 6F 72 61 20 64 65 20 61 63 6F 72 64 61 72
6D 65 20 64 65 20 66 65 63 68 61 73 20 65 73 70 65 63 69 66 69 63 61 73 20 6F 20 64 65 20
64 6F 6E 64 65 20 64 65 6A 6F 20 6C 61 73 20 63 6F 73 61 73 20 2C 20 70 65 72 6F 20 6E
6F 20 6C 6F 20 64 75 64 65 73 2C 20 73 69 65 6D 70 72 65 20 72 65 63 75 65 72 64 6F 20
74 75 20 73 6F 6E 72 69 73 61 20 65 6E 20 6C 6F 73 20 6D 65 6A 6F 72 65 73 20 79 20 70
65 6F 72 65 73 20 6D 6F 6D 65 6E 74 6F 73 2E 0A 0A 54 65 20 6D 65 72 65 63 65 73 20
6C 6F 20 6D 65 6A 6F 72 20 64 65 20 65 73 74 65 20 6D 75 6E 64 6F 20 61 73 69 20 63
6F 6D 6F 20 6D 69 6C 65 73 20 64 65 20 61 62 72 61 7A 6F 73 20 79 20 62 65 73 6F 73 2E 0A

Dedicado a mi familia, la cual ha tenido que convivir durante dos meses con tres antenas así como una estación meteorológica y estar pendientes en mi ausencia de posibles cortes de luz y de no mover ningún cable para evitar fallos en los dispositivos.

Capítulo 1

Introducción

En 1907, un joven inventor italiano llamado Guglielmo Marconi consiguió poder transmitir información desde un dispositivo emisor hasta un receptor mediante ondas de radio. Desde entonces, el transcurso de la humanidad cambió de manera drástica, ya que permitió que la comunicación fuera no solo fuera instantánea, si no que tuviera un alcance nunca antes visto y además no necesitaba de una infraestructura clásica como era el caso del telégrafo.

A partir de ese momento, las aplicaciones de las ondas de radio no han parado de crecer, siendo unos de los ejemplos más claros:

- El uso de la radio como medio de información
- El desarrollo de las comunicaciones móviles que nos mantienen comunicados en todo momento
- La transmisión de información entre dispositivos IoT (música, imágenes, etc...)

1.1. División institucional de las franjas de frecuencias electromagnéticas

Debido a la naturaleza de propagación de las ondas electromagnéticas y de las propiedades de la misma que pueden usarse para transmitir información, fue necesario dividir el

espectro electromagnético en función de las necesidades técnico-sociales con el fin de evitar la saturación de señal en ciertas partes del espectro. En España (lugar donde se realiza este estudio), la división del espectro queda recogido en el Boletín Oficial del Estado de la siguiente forma:

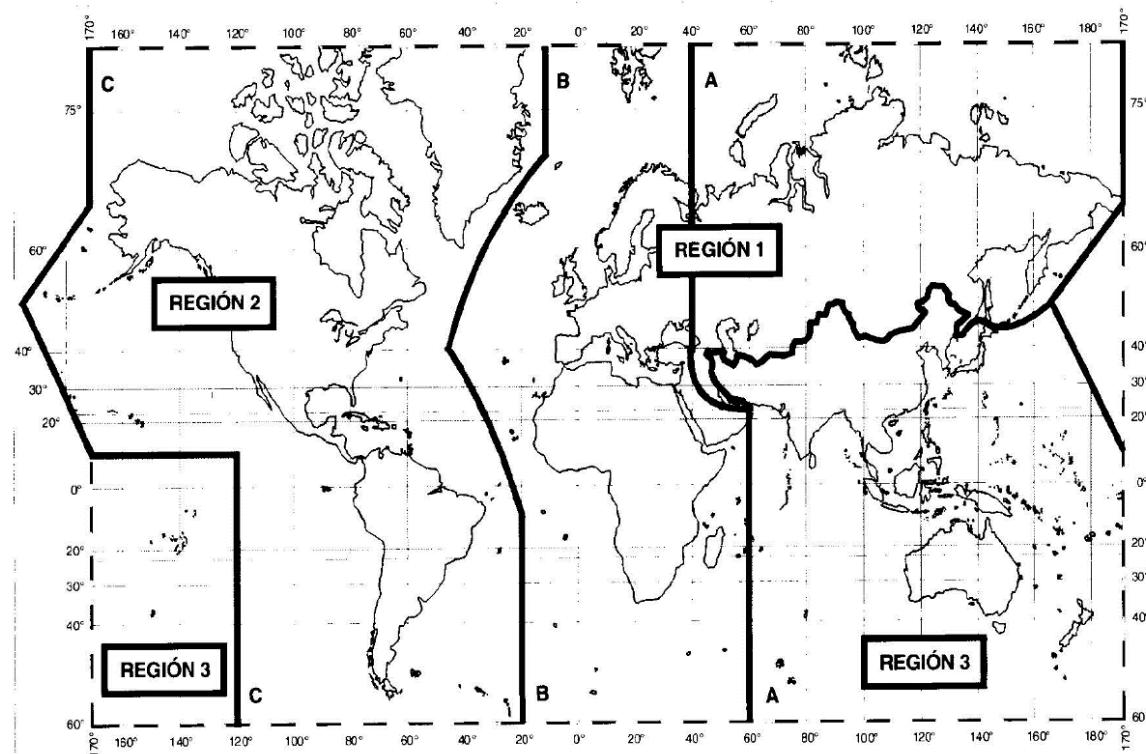


Figura 1.1: División del espectro electromagnético en las bandas de estudio de este trabajo. Extraído de Orden ETD/1449/2021, de 16 de diciembre, por la que se aprueba el Cuadro Nacional de Atribución de Frecuencias⁶

ATRIBUCIÓN A LOS SERVICIOS según el RR de la UIT		
410 - 460 MHz		
Región 1	Región 2	Región 3
410 - 420	FIJO MÓVIL, salvo móvil aeronáutico INVESTIGACIÓN ESPACIAL (espacio-espacio) 5.268	
420 - 430	FIJO MÓVIL, salvo móvil aeronáutico Radiolocalización 5.269 5.270 5.271	
430 - 432 AFICIONADOS RADIOLOCALIZA- CIÓN 5.271 5.274 5.275 5.276 5.277	430 - 432 RADIOLOCALIZACIÓN Aficionados 5.271 5.276 5.278 5.279	
432 - 438 AFICIONADOS RADIOLOCALIZA- CIÓN Exploración de la Tierra por satélite (activo) 5.279A 5.138 5.271 5.276 5.277 5.280 5.281 5.282	432 - 438 RADIOLOCALIZACIÓN Aficionados Exploración de la Tierra por satélite (activo) 5.279A 5.271 5.276 5.278 5.279 5.281 5.282	
438 - 440 AFICIONADOS RADIOLOCALIZA- CIÓN 5.271 5.274 5.275 5.276 5.277 5.283	438 - 440 RADIOLOCALIZACIÓN Aficionados 5.271 5.276 5.278 5.279	

ATRIBUCIÓN NACIONAL USOS OBSERVACIONES		
410 - 460 MHz		
410 - 420 FIJO MÓVIL, salvo móvil aeronáutico INVESTIGACIÓN ESPACIAL (espacio-espacio)	M M P	5.268 UN-31 UN-74 UN-77 UN-154, UN-156
420 - 430 FIJO MÓVIL, salvo móvil aeronáutico Radiolocalización	M M M	UN-31 UN-74 UN-97 UN-154, UN-156
430 - 432 AFICIONADOS RADIOLOCALIZACIÓN	E M	
432 - 438 AFICIONADOS RADIOLOCALIZACIÓN Exploración de la Tierra por satélite (activo)	* M M	5.138 5.279A Banda de aplicaciones ICM 433.05-434.79 MHz UN-30, UN-32 UN-115, UN-154
438 - 440 AFICIONADOS RADIOLOCALIZACIÓN	E M	* Usos E y C (según notas UN)

ATRIBUCIÓN A LOS SERVICIOS según el RR de la UIT		
890 - 1300 MHz		
Región 1	Región 2	Región 3
890 - 942 FIJO MÓVIL, salvo móvil aeronáutico 5.317A RADIODIFUSIÓN 5.322 Radiolocalización	890 - 902 FIJO MÓVIL, salvo móvil aeronáutico 5.317A Radiolocalización 5.318 5.325	890 - 942 FIJO MÓVIL, salvo móvil aeronáutico Radiodifusión 5.317A 5.322
	902 - 928 FIJO Aficionados Móvil, salvo móvil aeronáutico 5.325A Radiolocalización 5.150 5.325 5.326	
	928 - 942 FIJO MÓVIL, salvo móvil aeronáutico 5.317A Radiolocalización	5.327
5.323	5.325	
942 - 960 FIJO MÓVIL, salvo móvil aeronáutico 5.317A RADIODIFUSIÓN 5.322	942 - 960 FIJO MÓVIL 5.317A	942 - 960 FIJO MÓVIL, salvo móvil aeronáutico
5.323		5.320
960 - 1164 MÓVIL AERONAUTICO (R) 5.327A RADIONAVEGACIÓN AERONÁUTICA 5.328 5.328AA		960 - 1164 MÓVIL AERONAUTICO (R) RADIONAVEGACIÓN AERONÁUTICA
		5.327A 5.328 5.328AA UN-154

ATRIBUCIÓN NACIONAL USOS OBSERVACIONES		
890 - 1300 MHz		
890 - 942 FIJO MÓVIL, salvo móvil aeronáutico Radiodifusión	M P R	5.317A 5.322 Sistemas terrenales capaces de prestar servicios de comunicaciones electrónicas (890-915/925-942 MHz) UN-40, UN-41 UN-104 CT1-E UN-154
942 - 960 FIJO MÓVIL, salvo móvil aeronáutico	M P	5.317A 5.322 Sistemas terrenales capaces de prestar servicios de comunicaciones electrónicas (942-960 MHz) UN-41, UN-154
960 - 1164 MÓVIL AERONAUTICO (R) 5.327A RADIONAVEGACIÓN AERONÁUTICA 5.328 5.328AA	R R	5.327A 5.328 5.328AA UN-154

Figura 1.2: División del espectro electromagnético en las bandas de estudio de este trabajo. Extraído de Orden ETD/1449/2021, de 16 de diciembre, por la que se aprueba el Cuadro Nacional de Atribución de Frecuencias⁶

Para este trabajo nos centraremos en las frecuencias de 433MHz, 930MHz y 1090MHz, que a la vista de las anteriores imágenes se usan para emitir información de dispositivos, sensores, etc... (433MHz), la comunicación entre teléfonos (930MHz) y la señal de los transpondedores de los aviones (1090MHz). Además, todas estas emisiones de radiofrecuencias están presentes en cualquier entorno donde habite el ser humano, por lo que actuarían a modo de indicadores de la presencia del mismo y por ende de la posible contaminación.

1.1.1. 433 MHz

La frecuencia de 433MHz está exenta de un impuesto por emisión de señales, por lo que muchos fabricantes de sensores optan por emplear esta franja para que sus dispositivos puedan emitir y recibir información.

Algunos ejemplos de estos sensores pueden ser: las estaciones meteorológicas de uso doméstico, sensores de humo, etc. Pero uno de los sensores más curiosos y que podrían tener un buen desempeño en este trabajo serían los sensores TPM de las ruedas de los coches.

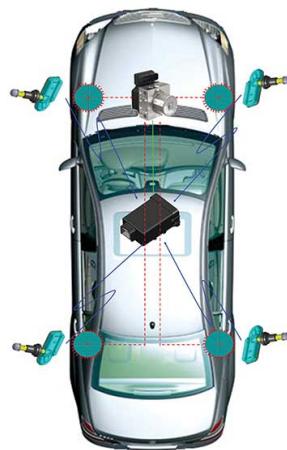


Figura 1.3: Diagrama de los sensores TPM en los coches

La importancia de estos sensores viene dada por la captura y decodificación de la temperatura y presión de las ruedas próximas a la antena y por tanto la captura y posterior

identificación de la presencia de vehículos que transitan por la zona. Dicho en otras palabras, permite contabilizar el número de vehículos que circulan en un radio cercano a la antena.

Estos sensores pueden ser capturados mediante el script `rtl_433` del usuario de Github [merbanan](#).

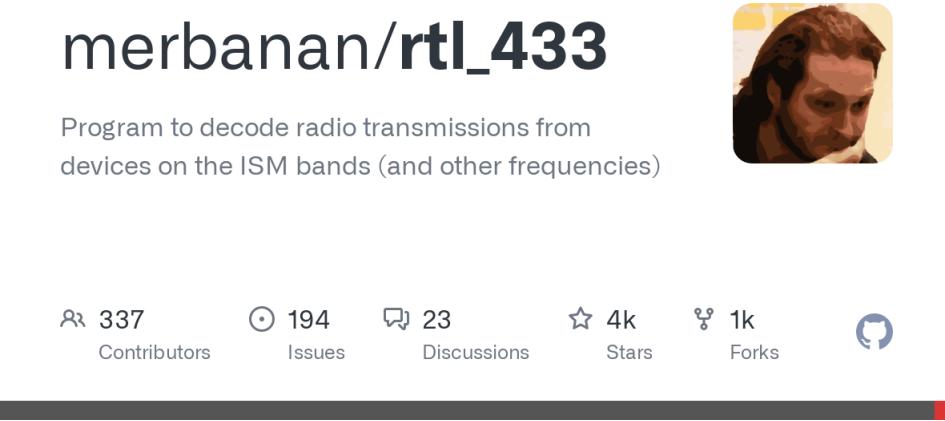


Figura 1.4: Ficha de `rtl43310` en Github

Pero este sistema tiene varios inconvenientes, siendo estos:

- No todos los vehículos tienen implementados estos sensores, ya que la normativa que obliga a equipar dichos sensores entró en vigor en Noviembre de 2014
- El radio de captura de información es de 10 metros pudiendo llegar a 40 metros si se amplifica la señal como bien se explica en el artículo *Security and privacy vulnerabilities of in-car wireless networks: a tire pressure monitoring system case study*¹³

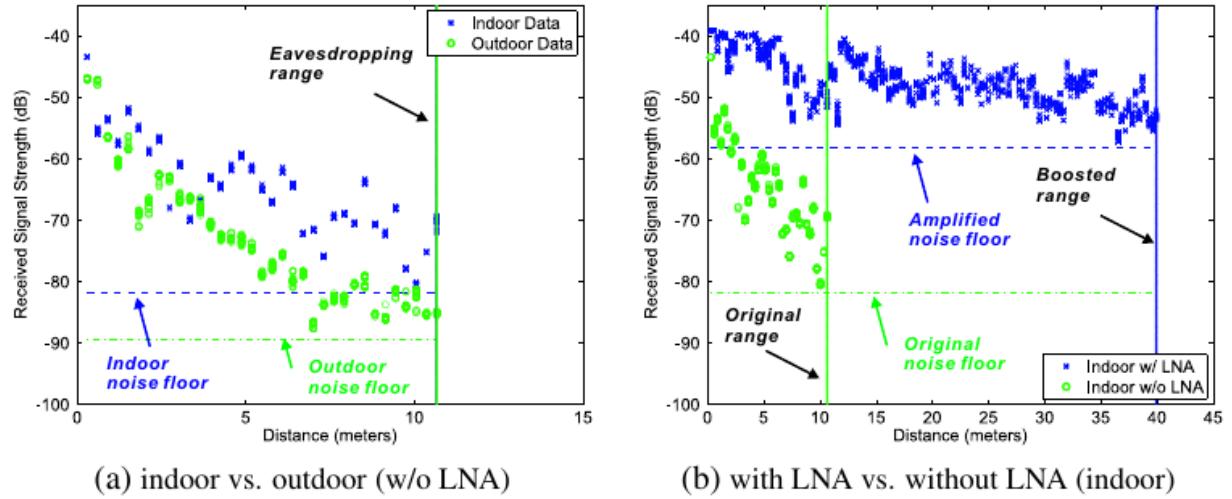


Figura 1.5: Distancia de captura de datos de los sensores TPMS. Extraído de *Security and privacy vulnerabilities of in-car wireless networks: a tire pressure monitoring system case study*¹³

1.1.2. 1090 MHz

Al contrario de la frecuencia 433MHz, la 1090MHz está fuertemente regulada, ya que en ella emiten de forma constante los transpondedores ¹ de los aviones. De esta forma se puede rastrear y seguir a todas las aeronaves en un radio que depende de la calidad de la antena empleada. De modo que se podrían contar el número de aviones que sobrevuelan la zona de estudio.

¹Dispositivos de emisión de señales activa que se emplean para emitir la posición actual de una aeronave durante el vuelo

Para obtener dicha información, se tendría que emplear el script dump1090 del usuario de Github antirez.

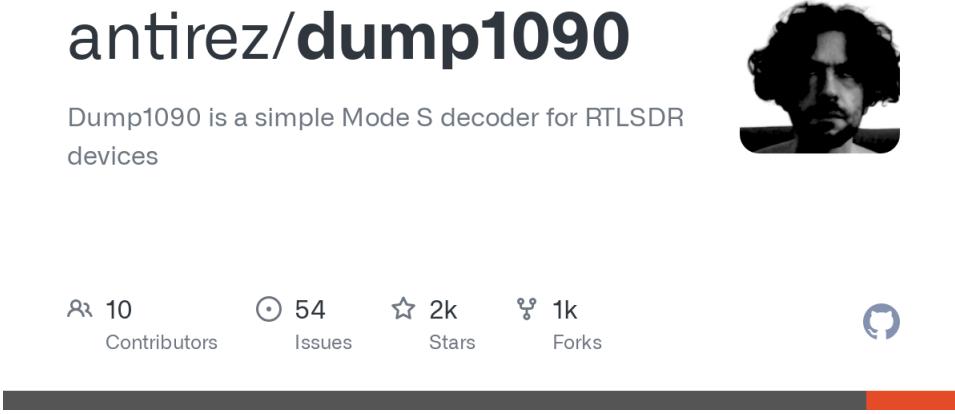


Figura 1.6: Ficha de *dump1090*¹ en Github

El único inconveniente de la recogida de datos en el presente estudio es la importancia que tiene Las Rozas de Madrid, lugar en el que se ha realizado el estudio, la cual es una zona de exclusión de aviones y no nos aportarían datos suficientes.

1.1.3. 930 MHz

Finalmente, si ponemos el foco en las franjas de los 930MHz encontramos todo el ecosistema de las telecomunicaciones que emplean las torres telefónicas como los dispositivos móviles para comunicarse entre ellos. Con esta información se puede estimar el número de vehículos que pasan por una zona concreta a partir de las conexiones de los teléfonos a las diferentes antenas telefónicas. A modo de ejemplo, los Cuerpos de Seguridad del Estado pueden emplear *IMSI catches* activos para saber si un delincuente está por la zona sin tener que pedir los registros a las teleoperadoras (estos casos son más comunes en países como Estados Unidos)

Es en este punto donde se abren infinidad de posibilidades de estudios, desde afluencia de personas extranjeras (turismo), detección de horarios de ciertos teléfonos, triangulación de teléfonos/personas, detección de tráfico, etc.

Para este estudio, nos centraremos únicamente en la afluencia de personas en partes claves de Las Rozas. En la siguiente imagen se puede apreciar un mapa temático de Las Rozas, donde los puntos negros representan las antenas de estudio de este trabajo.

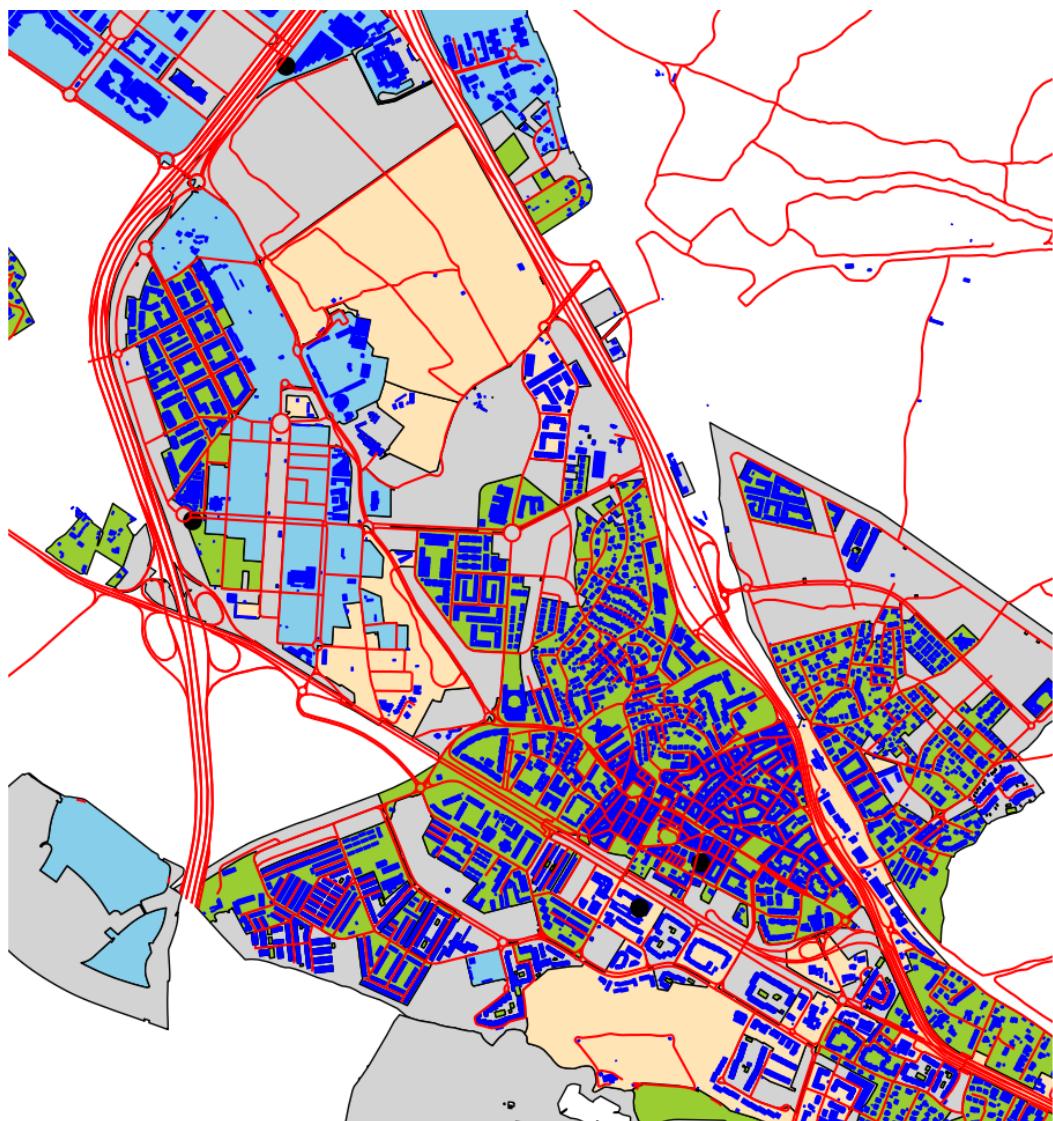


Figura 1.7: Mapa temático de Las Rozas de Madrid. Elaboración propia con los datos del IGNE.

Para capturar las comunicaciones IMSIS de dichas antenas, se emplearan 3 RTL-SDR² a modo de *IMSI catchers pasivos* haciendo uso del paquete *grgsm* y del script *IMSI-catcher* del usuario de Github Oros24.

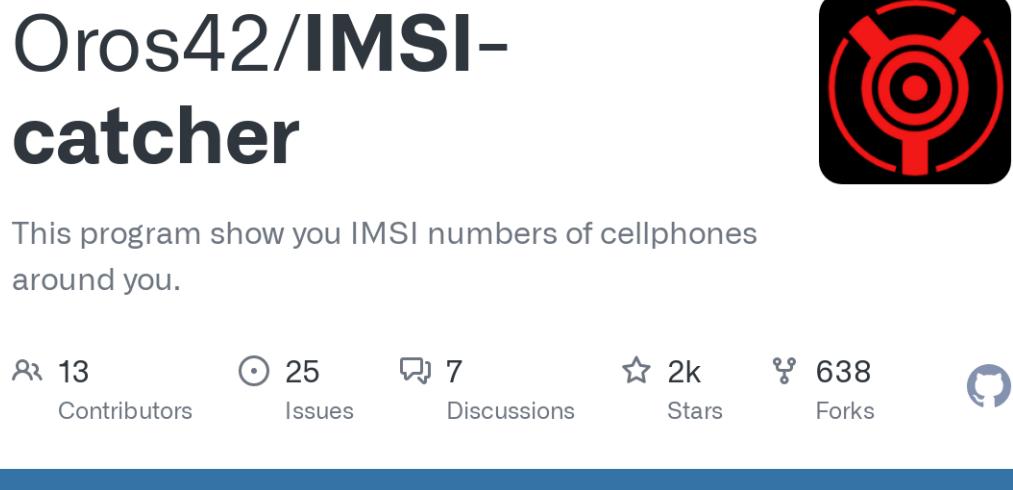


Figura 1.8: Ficha de *IMSI-catcher*¹² en Github

De esta forma, se monitorizarán y registrarán constantemente los identificadores IMSI (International Mobile Subscriber Identity) el cual es un identificador único asignado a cada tarjeta SIM.



Figura 1.9: Estructura de un numero IMSI

- **MCC:** Mobile Country Code
- **MNC:** Mobile Network Code
- **MSIN:** Mobile Subscription Identification Number

²Dispositivo usb que permite escanear una pequeña parte del espectro de radio frecuencias y retornar dicha información a un software específico que permite la posterior decodificación y análisis de las ondas

1.2. IMSI catchers

Una vez explicadas las diferentes bandas desde donde se pueden extraer información, hablaremos sobre el dispositivo empleado para capturar dicha información, siendo estos los *IMSI catchers*, los cuales son capaces de capturar las conexiones telefónicas ³ que ocurren en una o varias antenas. Dicha captura puede ser obtenida por dos vías:

- **De forma activa:** el *IMSI catcher* se hace pasar por una antena real forzando a que todas las conexiones pasen por él, para que posteriormente sea redirigida a una antena verdadera de forma que el usuario no sospeche nada.
- **De forma pasiva:** el *IMSI catcher* se limita únicamente a escuchar el inicio de la conexión entre un teléfono y una antena real captando el inicio del protocolo de las comunicaciones de identificación que realizan ambos dispositivos

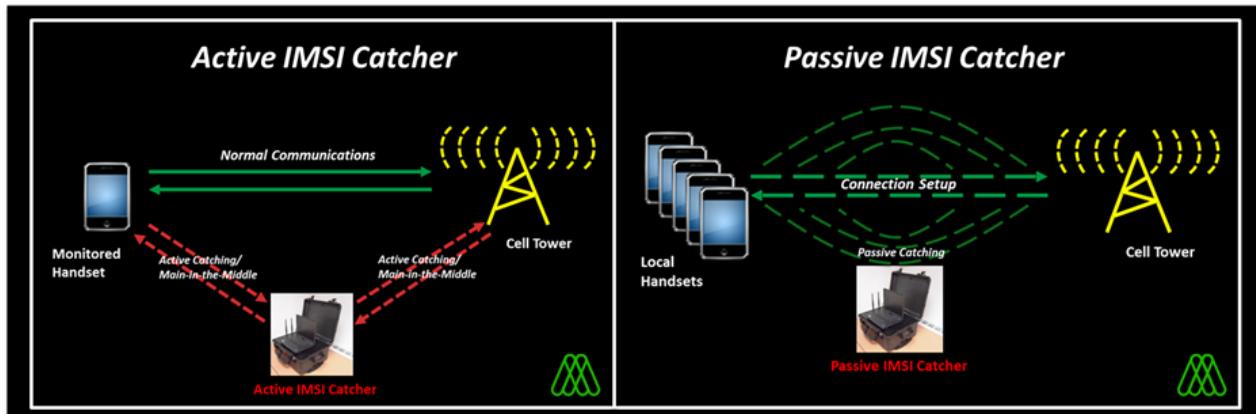


Figura 1.10: Diferencias entre un *IMSI catcher* activo y pasivo

Para este trabajo se han empleado los SDR-RTL a modo de *IMSI catchers* pasivos. En el anexo de bases legales se trata sobre la legalidad de este tipo de práctica.

³En el apartado de modelos se denotara como T_t (T de teléfonos) a la suma de teléfonos presentes en el instante t, siendo t una hora cualquiera

1.3. Estación meteorológica

Finalmente, pasamos a describir brevemente la forma en la que se han recopilado los datos ambientales de una pequeña parte de Las Rozas de Madrid. Para ello, se ha optado por construir y programar una estación meteorológica, ya que no existe una alternativa de código y datos libres en el mercado que cubriera las necesidades de este trabajo. Esta decisión ha permitido ir mejorando la estación a lo largo del tiempo y adaptarla a las circunstancias, por no decir que el proceso de creación de la misma ha resultado ser una experiencia muy enriquecedora a nivel de conocimientos electrónicos.

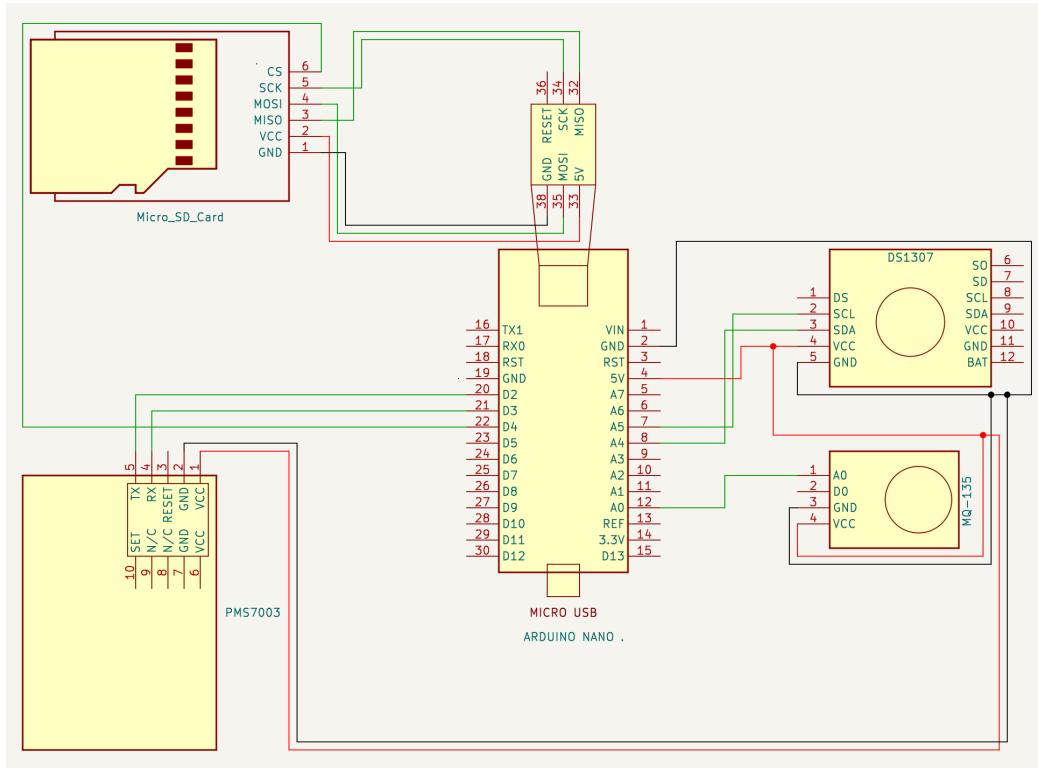


Figura 1.11: Diagrama de conexiones y componentes de la estación meteorológica

A continuación, se describen los sensores empleados en la estación meteorológica la cual se ha construido manualmente:

1.3.1. PMS7003

Este sensor es el encargado de recopilar la información sobre las partículas en suspensión⁴, siendo capaz de detectar elementos presentes en el ambiente cuyo rango de tamaño puede oscilar desde los 2.5 micrómetros hasta los 100 micrómetros.

- **PM 2.5:** partículas menores a 2.5 micrómetros que suelen ser el resultado de la presencia de industria pesada, quema de carbón, combustiones, vehículos, etc... Al ser tan pequeñas pueden estar suspendidas en el aire varios días (dependiendo de la climatología: lluvias, vientos, etc...), además se pueden introducir en el sistema respiratorio muy fácilmente llegando a causar ciertos problemas siempre que la concentración de este tipo de partículas en el aire sea muy alta.
- **PM 10:** partículas comprendidas entre 2.5 y 10 micrómetros que suelen ser partículas de polvo o residuos generados por la fricción de materiales. A diferencia de las partículas PM 2.5, estas suelen estar suspendidas en el aire varias horas y no causan tantos problemas en el organismo como las PM 2.5. Esto no quita que puedan suponer ciertas molestias al organismo tales como: irritación de ojos, garganta, nariz, etc...

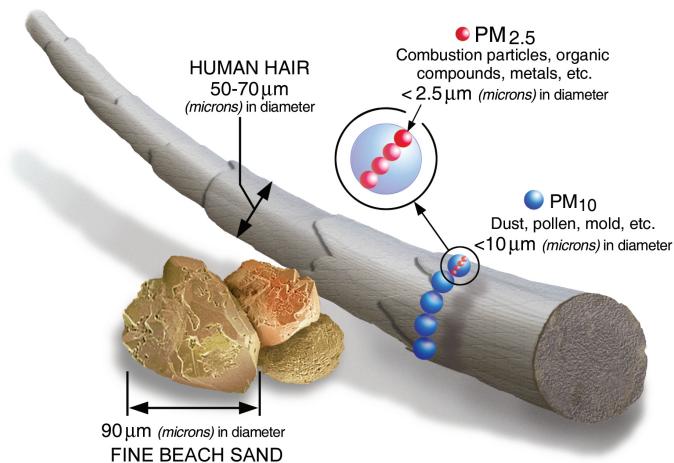


Figura 1.12: Escala dimensional de las partículas de estudio frente a un cabello humano

⁴En el apartado de modelos se denotara como C_t (C de contaminación) a la media de partículas en suspensión en el instante t, siendo t una hora cualquiera

1.3.2. MQ 135

Este sensor es el encargado de recopilar la información sobre la concentración de varios tipos de gases en el aire. A diferencia del sensor PMS7003, que manda la información ya procesada, el sensor MQ 135 manda una señal analógica referente al voltaje que pasa por la resistencia que reacciona químicamente al ambiente. De esta forma, a mayor valor del voltaje, mayor será la presencia de gases en el ambiente.

Para este trabajo, se ha tomado como medida para detectar la presencia de gases el voltaje de este dispositivo. No obstante, se puede obtener una relación entre el voltaje y la siguiente imagen con el fin de medir la presencia de los gases, pero como ya se ha comentado, el sensor no es capaz de distinguir entre dos gases, simplemente determinar si uno o varios de los elementos esta presente o no.

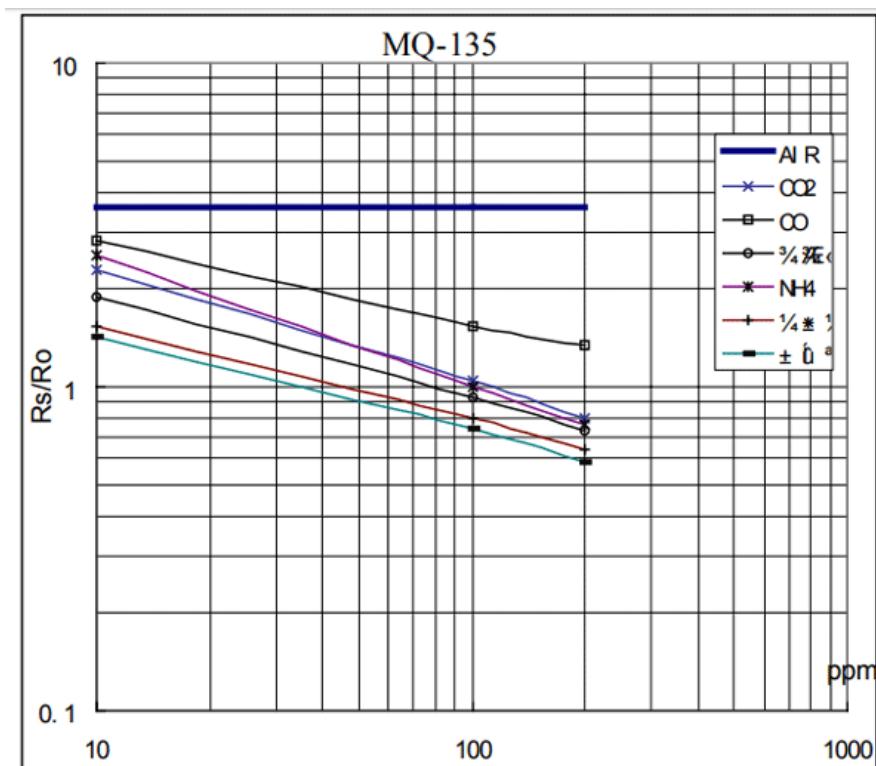


Figura 1.13: Log chart asociado al sensor MQ 135. Extraído de Technical data. MQ 135 Gas sensor¹¹

Capítulo 2

Objetivos

El principal objetivo de este trabajo es el de intentar predecir la contaminación ambiental en las próximas horas mediante la información emitida por diversos dispositivos presentes en nuestro día a día, tales como teléfonos, vehículos o aviones así como los anteriores mediciones de la contaminación ambiental.

Además, todo esto este trabajo ha realizado con la filosofía de *open data* y código abierto, de forma que cualquier persona pueda replicar este trabajo sin necesidad de que se tenga que recurrir a software o herramientas *close source* o privativas.

El segundo objetivo de este trabajo es el de dar visibilidad a los datos invisibles de las radiofrecuencias que nos rodean y que pueden resultar muy útiles para nuestro día a día además de enriquecedores.

Por último, y no menos importante, está el objetivo de aplicar todo lo aprendido en el Máster de Minería de Datos así como en el Grado de Estadística Aplicada en el mundo real con toda la problemática que ello conlleva (bases de datos sin depurar, errores de lecturas de datos, sensores que no funcionan como deberían, etc...)

Capítulo 3

Metodología

3.1. Recopilación de datos

La recopilación de datos se ha realizado durante los meses de septiembre y octubre (24-9-2022 hasta el 30-10-2022), monitorizando y recopilando sin interrupción y de forma masiva las conexiones telefónicas de Las Rozas de Madrid en 3 ubicaciones diferentes así como la calidad del aire en una zona intermedia resultante de la triangulación de las antenas.

A lo largo de dicha ventana temporal se han obtenido un total de 915909 conexiones telefónicas de IMSIS, de las cuales 450666 son únicas, es decir que existen IMSIS asociados a personas que transitan por Las Rozas de Madrid con frecuencia. En lo que a la estación meteorológica respecta, se han recopilado aproximadamente 300000 registros ambientales con un intervalo de recogida de datos de 10 segundos.

3.1.1. Ejemplos datos IMSIS

A continuación se muestran ejemplos de los datos en brutos de los IMSIS¹:

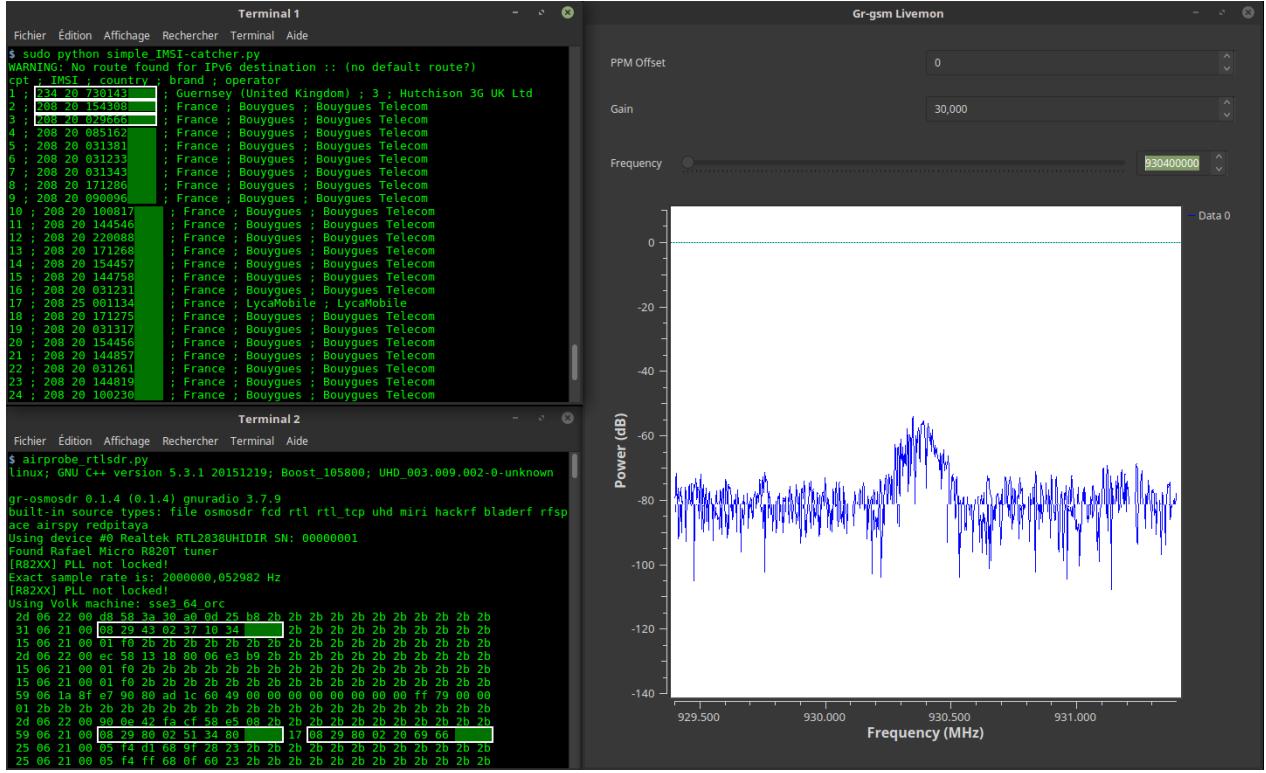


Figura 3.1: Ejemplo de salidas de datos del imsi catcher pasivo del repositorio de Oros42¹²

3.1.2. Ejemplos datos ambientales

```
1/10/22 0:0:9#10#11#13#10#11#13#1764#520#32#4#2#0 4.91
1/10/22 0:0:19#9#11#12#9#11#12#1713#513#32#5#3#0 4.91
1/10/22 0:0:30#10#12#12#10#12#12#2004#534#38#7#0#0 4.96
1/10/22 0:0:40#10#12#12#10#12#12#2004#534#38#7#0#0 4.81
1/10/22 0:0:50#8#9#8#9#9#1488#432#39#0#0#0 4.95
1/10/22 0:1:0#8#10#10#8#10#1494#436#43#0#0#0 4.90
```

¹Por motivos legales y de privacidad, se muestra el ejemplo de datos proporcionado en el github de Oros42, ya que si se mostraran los datos de imsí de Las Rozas se estaría exponiendo información de teléfonos ajenos que podrían usarse con fines distintos a los de este trabajo. Para mas información, leer las bases legales de este trabajo

3.1.3. Dificultades en la recopilación de datos

Como en cualquier aplicación del análisis y recogida de datos en problemas reales se pueden dar errores en la recogida de los datos, siendo las dificultades más recurrentes en este trabajo:

- **Problemas con el suministro eléctrico:** los dispositivos empleados para este trabajo necesitan de una fuente eléctrica constante, por lo que cualquier subida de tensión, apagón, cables de alimentación sueltos, etc... causa que se apaguen y se deje de recopilar los datos. En algunas ocasiones se dieron casos donde se produjeron problemas eléctricos a altas horas de la noche o cuando no había nadie cerca para activar todo de nuevo.
- **Problemas de estabilidad del código:** los dispositivos dejan de funcionar ante cualquier fallo de código o bug.
- **Interferencias de señales:** la información no se captura debidamente ante la presencia de interferencias en las señales.

3.2. Bases de datos

Debido a la naturaleza temporal de los datos, la estructura de las mismas están orientadas a bases de datos temporales, con el inconveniente de que es necesario cuadrar las escalas temporales de los datos de radiofrecuencias y los meteorológicos

3.3. Agregación de los datos

Para poder realizar este estudio, es necesario construir y agregar los datos en una única base de datos, ya que los IMSIS se recopilan a medida que se realiza una conexión mientras que la estación meteorológica recopila la información cada 10 segundos.

Para solucionar esto se ha optado por sumar todas las ocurrencias de los IMSIS por horas, mientras que en el caso de los datos de la estación meteorológica, se ha calculado la media aritmética de los datos por horas. Una vez calculadas dichas agregaciones, se procedió a la unión de los datos en una única base de datos.

Una vez agrupados los datos se analizaran las series temporales para detectar posibles fallos en la recopilación de los mismos. La forma en la que se procederá será la siguiente:

- Si los datos muestran un deterioro severo e incorregible, se procederá con la eliminación de la información.
- Si los datos muestran un deterioro moderado, se intentara reconstruir la información con la agregación de los mismos con el fin de intentar completar y suavizar los errores.

Esto da como resultado una base de datos con la que se puede trabajar a con series temporales e incluso con modelos clásicos y o alternativos. Por otra parte, en el caso de los datos ambientales se obtiene como efecto secundario una suavización de datos y evita en la medida de lo posible la aparición de outliers.

A continuación se presenta un extracto de la base de datos agregada con la que se ha trabajado²:

	particles_100um (C)	Total_imsis (T)	Viento_fuerte (V)
0	0.14	296.00	0
1	0.11	182.00	0
2	0.05	141.00	0
...
595	0.20	884.00	0
596	0.20	464.00	0
597	0.28	258.00	0

²Esta base de datos consta únicamente de tres variables ya que durante la fase de recopilación de datos ocurrieron eventos que afectaron a la calidad del dato. De forma que de las 11 variables originales, 4 sufrieron errores irreparables, 3 tenían un comportamiento idéntico que podría causar problemas de autocorrelacion y finalmente otras 3 variables se agregaron en una única variable con el fin de completar y corregir los errores de los datos

De forma que los datos y sus unidades son:

- **Particles_100um**: número medio de partículas de más de 100 micrómetros en 0.1 Litros de aire por hora y será representado como C de contaminación en las ecuaciones de más adelante.
- **Total_imsis**: suma total de todos los IMSIS avistados en las antenas de estudio por hora y será representado como T de teléfonos en las ecuaciones de más adelante.
- **Viento**: variable booleana que indica si existió viento fuerte o no por hora y será representado como V de viento en las ecuaciones de mas adelante.

3.4. Limpieza y división de la base de datos

Debido a algunos problemas y corrupción de datos se ha decidido trabajar únicamente con el histórico de mes de octubre. En lo que a la división de la base de datos respecta, se ha dividido el histórico de octubre en dos conjuntos de datos.

- **Train**: desde el inicio del mes de octubre hasta el día 25 de octubre
- **Test**: desde el día 26 de octubre hasta el último día de octubre

3.5. Metodología SEMMA

Para este trabajo se empleará la metodología de trabajo SEMMA, que se caracteriza por ser una metodología de trabajo que se adapta y evoluciona a los resultados del análisis.

3.6. Selección de modelo

Debido a la naturaleza de los datos y del objetivo de este trabajo, la selección de modelos se reduce a la familia de series temporales. De esta forma, se emplearán dos modelos clásicos de series temporales y un modelo alternativo de redes neuronales.

En lo que a modelos clásicos de este trabajo respecta, usaremos un modelo multivariante de autorregresión vectorial con medias móviles (modelo VARMA) así como el modelo ARIMA donde sus ecuaciones teóricas quedarían de la forma:

Modelo VARMA³

$$C_t = \alpha + T_{t-1}\beta_1 + C_{t-1}\beta_1 + \dots + T_{t-p}\beta_{p-1} + C_{t-p}\beta_p + T_{t-1}\theta_1 + C_{t-1}\theta_2 + \dots + T_{t-p}\theta_p - 1 + C_{t-p}\theta_p + \epsilon_t \quad (3.1)$$

Modelo ARIMA

$$C_t = \alpha + V_t + T\beta_1 + C_{t-1}\beta_1 + \dots + C_{t-p}\beta_p + C_{t-1}\theta_2 + \dots + C_{t-p}\theta_p + \epsilon_t \quad (3.2)$$

Donde C es la contaminación, T los teléfonos presentes y p el retardo en horas.

Modelo LSTM

Las redes neuronales LSTM (Long short-term memory) se caracterizan por emplear conexiones de datos que permiten la retroalimentación de los mismos, pudiendo así emplear información del pasado con la cual afinar las estimaciones de los pesos. En la siguiente imagen se puede apreciar el esquema de como funciona este tipo de redes:

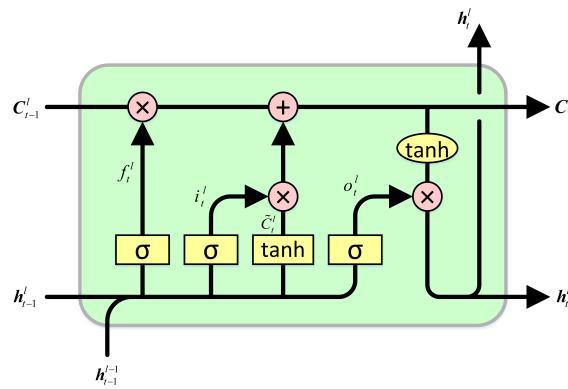


Figura 3.2: Esquema de la red neuronal LSTM

³En el modelo VARMA se omite la variable viento fuerte (V) por la complejidad del modelo así como los pocos registros de datos que podrían causar problemas de grados de libertades en las estimaciones de los parámetros

3.7. Metodología de selección de modelo

Para seleccionar los modelos clásicos, se buscarán todos los modelos posibles para cada combinación posible y nos quedaremos con el modelo con menor AIC y que tengan una diagnosis del modelo válidas. En el caso de la red neuronal, se buscará aquella combinación de parámetros que mejores resultados ofrezca. Una vez seleccionados, se compararan los modelos finalistas mediante el SEE

3.8. Software empleado en los modelos

Para la modelización de la familia de VARMA, se empleará el software estadístico statsmodels de python, concretamente la clase VARMAX, para el modelo ARIMA se empleara la clase ARIMA de statsmodels y finalmente para el modelo de redes neuronales se aplicará Keras.

Capítulo 4

Desarrollo

4.1. Descriptiva de los datos

En el siguiente gráfico se pueden observar la evolución de los IMSIS por hora, siendo los tres primeros gráficos los referentes a las antenas situadas en las tiendas Factory, Centro Comercial Burgo Centro y circuito de Karts Carlos Sainz. En dichos gráficos se aprecia como han existido problemas de pérdidas de señales o interferencias, es por eso, que la cuarta serie temporal hace referencia a la suma de las tres de arriba. Esto se hizo con el fin de descartar registros de la variable predictora.

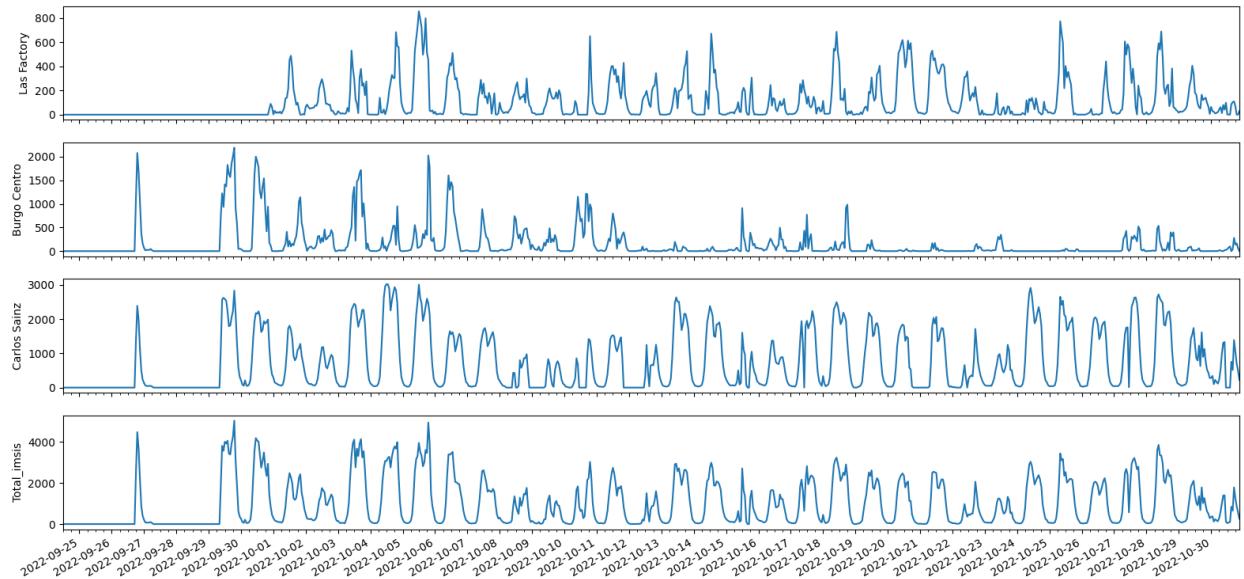


Figura 4.1: Series temporales de los datos entre los imsísculos detectados

En lo que al comportamiento de los IMSIS, se puede observar como existe un patrón temporal que concuerda con los horarios de las personas, es decir, a altas horas de la madrugada la presencia de IMSIS es muy baja, mientras que en las horas puntas se pueden apreciar los picos de movimientos para ir al trabajo o escuela y al regresar al domicilio.

En lo que respecta a los datos de la contaminación ambiental, podemos apreciar el siguiente gráfico donde se representa la evolución media de las partículas en suspensión así como la presencia de gases.

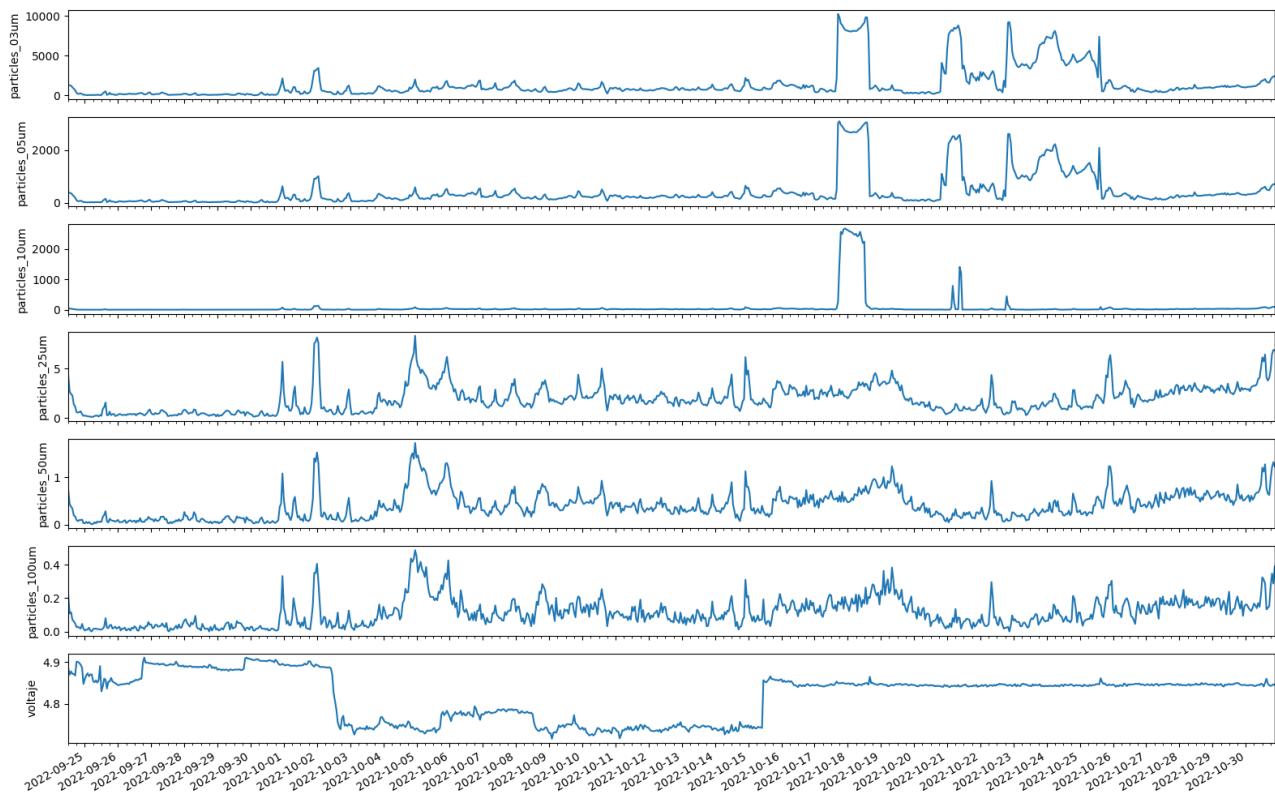


Figura 4.2: Series temporales de los datos de la estación meteorológica

A la vista del anterior gráfico, se puede apreciar claramente como la serie referente al voltaje de gases presentes y de partículas de 3 a 10 micrómetros presentan valores y comportamientos anómalos en comparación con las series de las partículas de más de 25 micrómetros. Es por ello, que descartaremos dichas series anómalas. Por otra parte, las series

de datos de las partículas tienen un comportamiento idéntico, de forma que con predecir una serie temporal se podría extraer dichos resultados a las otras series temporales.

Finalmente, se puede apreciar cambios de tendencias en la suspensión de partículas, es por ello que si nos fijamos en la evolución meteorológica en webs especializadas en temas ambientales se puede ver como dichos cambios de tendencias ocurren por el incremento de la velocidad del viento y las lluvias.

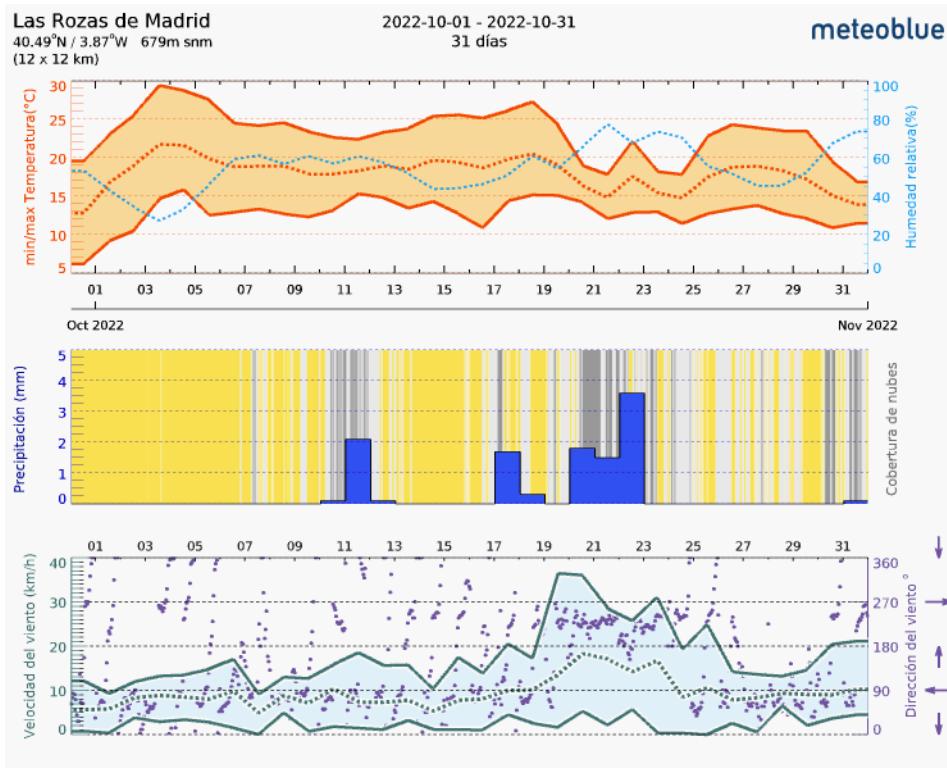


Figura 4.3: Datos meteorológicos de Las Rozas en Octubre. Extraído de meteoblue.com

Con todo lo anterior, podemos observar en los anteriores gráficos una posible relación entre teléfonos de la zona y la contaminación. Para contrastar dicha relación, se hará uso del contraste de causalidad de Grangers, pero para poder realizar el contraste de causalidad, primero se necesita que ambas series sean estacionales. Por lo que realizaremos el contraste

de Dickey-Fuller sobre las series de partículas de 100 micrómetros así como en la de teléfonos totales.

4.2. Estacionalidad de las series temporales

Para poder realizar cualquier modelo de series temporales, es necesario comprender el comportamiento de las mismas. ¿Tienen un comportamiento constante?, ¿crecen siempre?, ¿son ruidos blancos?, etc... Por tanto, empezaremos por lo básico y estudiaremos la estacionalidad de las series temporales. Para ello emplearemos el contraste de Dickey-Fuller, cuya hipótesis es:

$$H_0: \text{la serie no es estacionaria}$$

$$H_1: \text{la serie es estacionaria}$$

Para realizar el contraste de Dickey-Fuller, se empleará el contraste adfuller de statsmodels (Python). De esta forma, los p-valores asociados a las series de IMSIS y partículas respectivamente son de 0.01822 y 0.033214

En las cuatro series temporales se obtiene un p-valor por debajo del nivel de significación de 0.05 por lo que no existen evidencias estadísticas como para rechazar la hipótesis alternativa, esto es, que las series temporales son estacionarias y por tanto tienen una media y varianza constante.

4.3. Causalidad de Grangers de las series temporales

Una vez comprobado que las series son estacionarias, se procederá a contrastar el test de causalidad de Grangers, que se emplea para determinar si se puede usar los valores con cierto retardo de una serie temporal con el fin de poder predecir los valores de otra serie temporal, siendo el contraste de hipótesis asociado:

H_0 : no existe una relación entre las series temporales

H_1 : existe una relación entre las series temporales

Para realizar de causalidad de Grangers se empleara el contraste grangercausalitytests de statsmodels (Python) calculando de forma automática todos los retardos desde 1 hasta 24. De esta forma, se obtiene la siguiente tabla

Retardos	p-valor	Retardos	p-valor
1	0.04	13	0.00
2	0.00	14	0.00
3	0.00	15	0.00
4	0.00	16	0.00
5	0.00	17	0.00
6	0.00	18	0.00
7	0.00	19	0.00
8	0.00	20	0.00
9	0.00	21	0.00
10	0.00	22	0.00
11	0.00	23	0.00
12	0.00	24	0.00

Como se puede observar, para cualquier nivel de significación, no existen evidencias estadísticas como para rechazar la hipótesis nula, es decir, según los resultados del contraste, se aprecian indicios de que la serie temporal de IMSIS puede ser empleada para predecir la contaminación ambiental de la zona.

4.4. Ciclos de las series temporales

Una vez que se ha comprobado que las series temporales son estacionales y pueden usarse para predecir valores, debemos comprender mejor su comportamiento a lo largo del tiempo mediante sus ciclos. Para ello se estudiarán los ciclos temporales asociados a las series temporales de IMSIS y contaminación y emplearemos la técnica del periodograma, que se basa en la transformación rápida de Fourier (FFT) con el fin de detectar los ciclos

en los que la serie temporal se repite. Haciendo uso de la función periodograma disponible en la subclase signal de scipy.

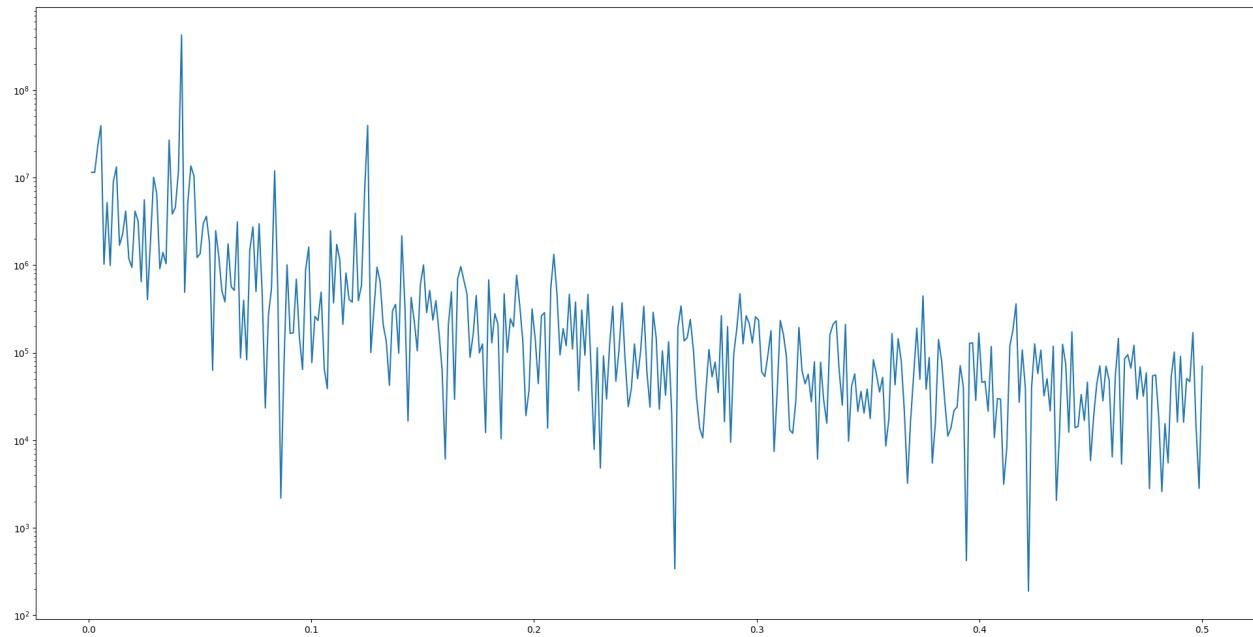


Figura 4.4: Periodograma de IMSIS

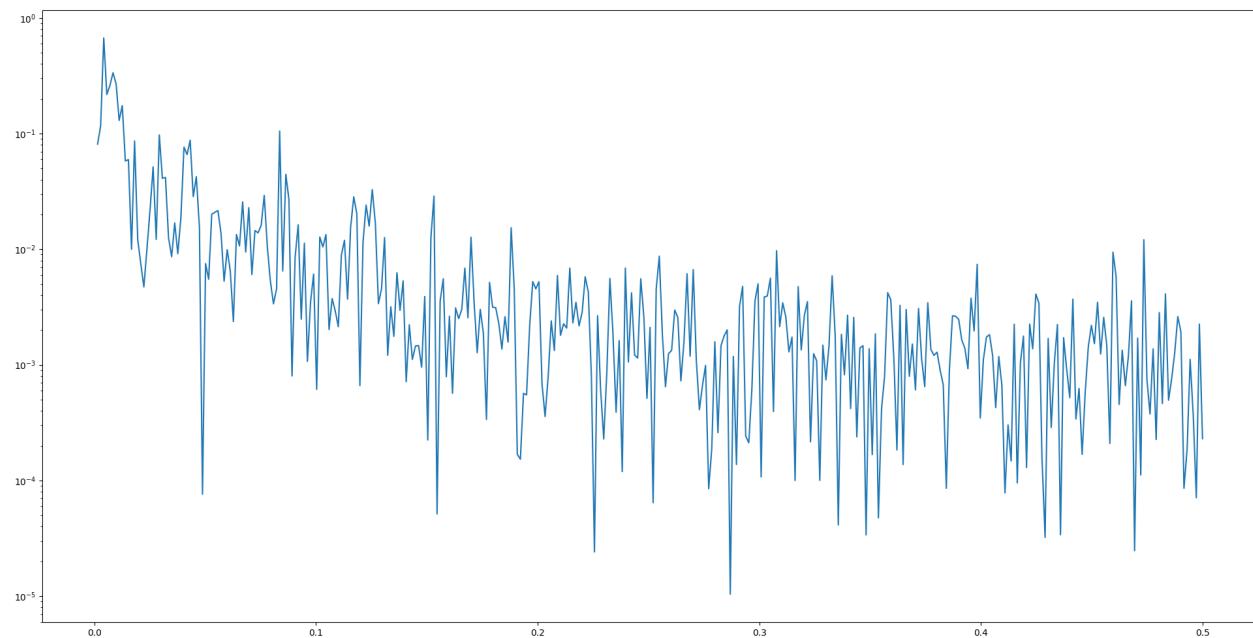


Figura 4.5: Periodograma de las partículas de 100 micrometros

Respecto a los valores máximos de los periodogramas, encontramos que en los IMSIS toma el valor de 425480622.2 (eje y), que está asociado al valor del eje x de 0.004178 y si dividimos 1 por dicho valor, obtenemos que los ciclos ocurren cada 24h. Si aplicamos esto mismo en la contaminación, obtenemos un valor del eje y de 0.670841, en el eje x un valor de 0.004178 dando un total de 240h entre cada ciclo.

Al tener ciclos muy dispares, resulta contraproducente descomponer ambas series temporales en estacionalidad y tendencia con el fin de emplear dichas descomposiciones para realizar las predicciones.

4.5. Modelado de series temporales ARIMA

Una vez ya se ha comprendido la naturaleza y comportamiento de ambas series temporales, debemos aplicar los modelos aptos para dichas condiciones, empezando por el modelo ARIMA.

Para determinar el modelo multivariante más óptimo, generaremos todas las combinaciones posibles de parámetros del modelo ARIMA y nos quedaremos con los modelos con menor AIC asociado.

	p	q	AIC	AICC	BIC	HQIC	MAE	MSE	SSE
0	0	0	1917.31	1917.38	1934.89	1924.15	0.06	0.01	3.87
1	0	1	-1604.72	-1604.62	-1582.75	-1596.17	0.05	0.00	2.20
2	0	2	-1715.42	-1715.28	-1689.06	-1705.16	0.04	0.00	1.79
3	0	3	-1802.65	-1802.46	-1771.90	-1790.68	0.04	0.00	1.52
4	0	4	-1854.79	-1854.55	-1819.64	-1841.11	0.04	0.00	1.39
5	1	0	-2004.43	-2004.33	-1982.47	-1995.88	0.03	0.00	1.20
6	1	1	-2012.35	-2012.21	-1985.99	-2002.09	0.03	0.00	1.18
7	1	2	-2010.32	-2010.13	-1979.56	-1998.34	0.03	0.00	1.18
8	1	3	-2009.27	-2009.02	-1974.12	-1995.58	0.03	0.00	1.18
9	1	4	-2007.43	-2007.13	-1967.89	-1992.04	0.03	0.00	1.18
10	2	0	-2012.15	-2012.01	-1985.79	-2001.89	0.03	0.00	1.18
11	2	1	-2010.24	-2010.05	-1979.48	-1998.26	0.03	0.00	1.18
12	2	2	-2008.42	-2008.18	-1973.28	-1994.74	0.03	0.00	1.18
13	2	3	-2007.42	-2007.12	-1967.88	-1992.03	0.03	0.00	1.18
14	2	4	-2005.43	-2005.06	-1961.50	-1988.33	0.03	0.00	1.18
15	3	0	-2010.24	-2010.05	-1979.48	-1998.26	0.03	0.00	1.18
16	3	1	-2008.19	-2007.94	-1973.04	-1994.50	0.03	0.00	1.18
17	3	2	-2008.92	-2008.61	-1969.38	-1993.52	0.03	0.00	1.18
18	3	3	-2006.44	-2006.06	-1962.50	-1989.33	0.03	0.00	1.18
19	3	4	-2004.63	-2004.18	-1956.30	-1985.81	0.03	0.00	1.18
20	4	0	-2008.81	-2008.57	-1973.67	-1995.13	0.03	0.00	1.18
21	4	1	-2007.00	-2006.69	-1967.46	-1991.60	0.03	0.00	1.18
22	4	2	-642.29	-641.91	-598.35	-625.18	0.06	0.01	3.87
23	4	3	-2005.42	-2004.97	-1957.09	-1986.60	0.03	0.00	1.18
24	4	4	-2002.43	-2001.90	-1949.71	-1981.90	0.03	0.00	1.18

Independientemente de seleccionar el modelo con menor AIC, es necesario reseñar la importancia de elegir correctamente el modelo adecuado, que estará condicionado no solo de una serie de factores técnicos (la significación de los parámetros, la validación de los residuos, etc.) si no por la existencia de otras métricas, que pese a representar lo mismo, tienen ciertos matices que conviene tener en cuenta a la hora de decidir el modelo.

Para evitar que este trabajo se extienda demasiado, se han seleccionado los mejor modelos de series temporales con menor AIC, que tengan parámetros significativos y que la validación de los supuestos sea favorable.

4.5.1. Modelado de ARIMA

A la vista de la anterior tabla, se puede apreciar como el mejor modelo disponible es el ARIMA(1, 0, 1). Este modelo trabaja bajo la premisa de que para predecir los valores futuros solamente se necesitan los valores anteriores de la serie temporal así como una pequeña corrección de una serie temporal auxiliar.

Para modelizar el modelo ARIMA, emplearemos el modelo ARIMA de statsmodels (Python) donde se realizara un modelo con los teléfonos totales y las partículas de 100 micrómetros.

Dep. Variable:	particles_100um	No. Observations:	598
Model:	ARIMA(1, 0, 1)	Log Likelihood	1012.177
Date:	lun, 21 nov 2022	AIC	-2012.355
Time:	11:11:13	BIC	-1985.993
Sample:	0 - 598	HQIC	-2002.091
Covariance Type:	opg		
		coef	std err
const	0.1251	0.016	7.801
Total_imsis	-1.094e-05	3.82e-06	-2.864
Viento_fuerte	-0.0285	0.015	-1.849
ar.L1	0.8738	0.020	44.675
ma.L1	-0.1394	0.041	-3.410
sigma2	0.0020	9.2e-05	21.796
Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	191.11
Prob(Q):	0.87	Prob(JB):	0.00
Heteroskedasticity (H):	0.79	Skew:	0.78
Prob(H) (two-sided):	0.10	Kurtosis:	5.29

Warnings:

- [1] Covariance matrix calculated using the outer product of gradients (complex-step).

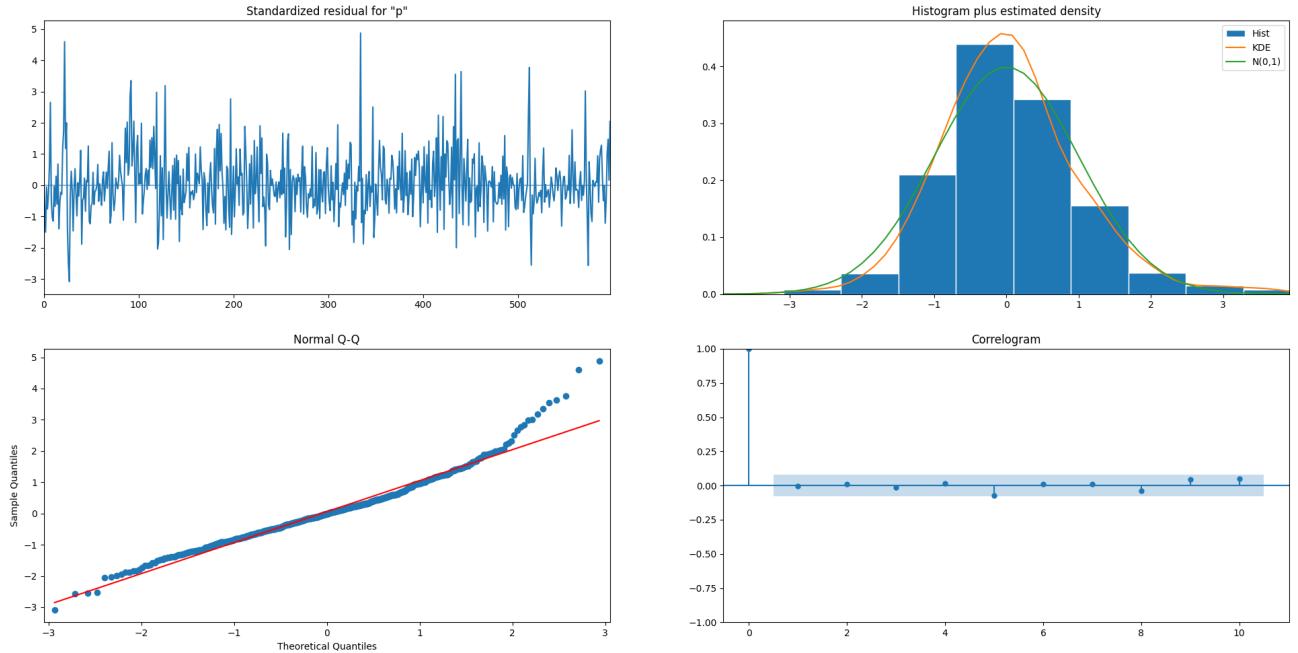


Figura 4.6: Diagnosis del modelo ARIMA

Como se puede observar en la salida del modelo, todos los parámetros son significativos para un nivel de significación menor a 0.10. Ademas, en lo que a la diagnosis del modelo respecta, los errores aparentan ser un ruido blanco y se asemejan a una distribución normal con una pequeña cola pesada a la derecha. Por otra parte, no aparenta sufrir de problemas de autocorrelación temporal. Por lo que a la vista de este modelo, parece indicar ser un buen candidato.

$$C_t = 0,1251 - 0,00001094T - 0,0285Viento fuerte + 0,8738C_{t-1} - 0,1394C\epsilon_{t-1} \quad (4.1)$$

Donde C es la contaminación y T los teléfonos presentes.

4.6. Modelado de series temporales multivariante

Para determinar el modelo multivariante más óptimo, generaremos todas las combinaciones posibles de parámetros del modelo VARMA y nos quedaremos con los modelos con menor AIC asociado.

p	q	AIC	AICC	BIC	HQIC	MAE	MSE	SSE
0	1	7852.80	7853.04	7887.95	7866.48	286.04	460256.35	2.752333e+08
0	2	8472.22	8472.76	8524.95	8492.75	314.44	548893.33	3.282382e+08
0	3	8417.84	8418.78	8488.14	8445.21	303.91	520478.02	3.112459e+08
0	4	7769.03	7770.49	7856.91	7803.25	242.76	363611.51	2.174397e+08
0	5	7725.02	7727.12	7830.47	7766.08	244.33	343143.87	2.052000e+08
0	6	7620.14	7622.99	7743.16	7668.03	231.41	324774.91	1.942154e+08
1	0	6981.42	6981.66	7016.57	6995.10	157.03	196793.34	1.176824e+08
2	0	6842.61	6843.15	6895.34	6863.14	136.58	160349.05	9.588873e+07
3	0	6846.36	6847.30	6916.66	6873.73	135.47	159278.98	9.524883e+07
4	0	6851.32	6852.78	6939.19	6885.53	135.27	158811.46	9.496925e+07
5	0	6856.80	6858.89	6962.25	6897.85	135.06	158686.45	9.489450e+07
6	0	6852.58	6855.43	6975.60	6900.47	134.25	157362.74	9.410292e+07
1	1	6879.68	6880.21	6932.40	6900.20	139.30	167746.63	1.003125e+08
1	2	6898.08	6899.01	6968.37	6925.45	138.60	170363.49	1.018774e+08
1	3	6903.06	6904.51	6990.93	6937.27	137.61	169343.84	1.012676e+08
2	1	6849.22	6850.15	6919.52	6876.59	136.67	160045.74	9.570735e+07
2	2	6856.40	6857.85	6944.27	6890.61	136.13	159911.54	9.562710e+07
2	3	6860.96	6863.06	6966.41	6902.02	134.69	159063.66	9.512007e+07
3	1	6854.40	6855.86	6942.28	6888.62	135.65	159290.72	9.525585e+07
3	2	6861.34	6863.44	6966.79	6902.40	135.73	159123.19	9.515567e+07
3	3	6868.83	6871.68	6991.85	6916.73	134.75	158916.69	9.503218e+07

4.6.1. Modelado de VAR

A la vista de la anterior tabla, vemos que el modelo con menor AIC asociado es el modelo VARMA(2, 0) el cual es lo mismo que un modelo VAR(2). Este modelo trabaja bajo la premisa de que para predecir un valor de la serie temporal se tiene que tener en cuenta los valores del pasado en todas las series temporales, es decir, que para poder predecir la contaminación ambiental, será necesario emplear la contaminación y los teléfonos con cierta cantidad de diferencias.

Para modelizar el modelo VAR, emplearemos el modelo VAR de statsmodels (Python) donde se realizará un modelo con los teléfonos totales y las partículas de 100 micrómetros.

Model:	VAR(2)	Log Likelihood	-3409.307
	+ intercept	AIC	6842.615
Date:	vie, 11 nov 2022	BIC	6895.338
Time:	22:13:10	HQIC	6863.142
Sample:	0 - 598		
Covariance Type:	opg		
Ljung-Box (L1) (Q):	0.01, 0.64	Jarque-Bera (JB):	179.63, 684.55
Prob(Q):	0.92, 0.43	Prob(JB):	0.00, 0.00
Heteroskedasticity (H):	0.81, 0.57	Skew:	0.78, 0.54
Prob(H) (two-sided):	0.14, 0.00	Kurtosis:	5.18, 8.13
		coef	std err
intercept	0.0137	0.004	3.338
L1.particles_100um	0.7153	0.036	19.797
L1.Total_imsis	-1.081e-05	3.97e-06	-2.725
L2.particles_100um	0.1329	0.033	3.993
L2.Total_imsis	1.573e-05	3.96e-06	3.969
		coef	std err
intercept	133.5998	9.35e-06	1.43e+07
L1.particles_100um	585.5205	8.98e-07	6.52e+08
L1.Total_imsis	1.3041	0.021	61.569
L2.particles_100um	-560.6333	1.69e-06	-3.32e+08
L2.Total_imsis	-0.4332	0.021	-20.533
		coef	std err
sigma2.particles_100um	0.0019	8.63e-05	22.168
sigma2.Total_imsis	1.609e+05	6.49e-08	2.48e+12
			P> z

Warnings:

- [1] Covariance matrix calculated using the outer product of gradients (complex-step).
- [2] Covariance matrix is singular or near-singular, with condition number 2.48e+27. Standard errors may be unstable.

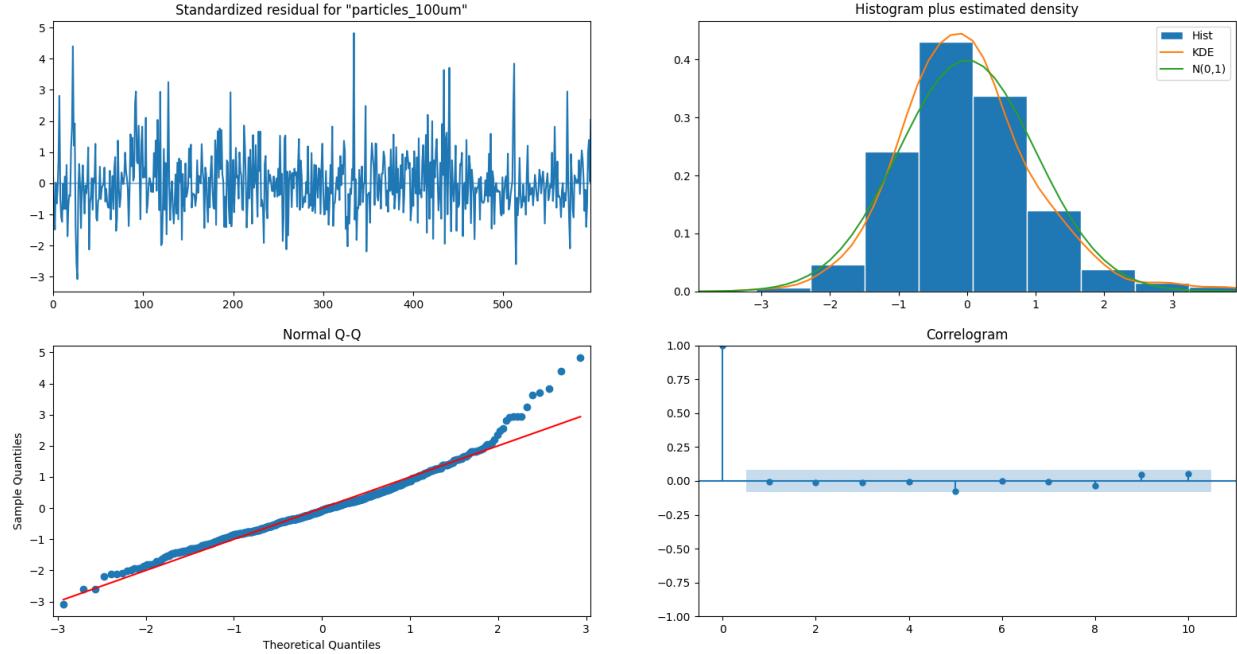


Figura 4.7: Diagnosis del modelo VARMA

Como se puede observar en la salida del modelo, todos los parámetros son significativos para cualquier nivel de significación. Además, en lo que a la diagnosis del modelo respecta, los errores aparentan ser un ruido blanco y se asemejan a una distribución normal con una pequeña cola pesada a la derecha. Por otra parte, no aparenta sufrir de problemas de autocorrelación temporal. Por lo tanto, parece indicar ser un buen candidato.

$$C_t = 0,0137 + 0,7153C_{t-1} - 0,00001081T_{t-1} + 0,1329C_{t-2} + 0,00001573T_{t-2} \quad (4.2)$$

Donde C es la contaminación y T los teléfonos presentes.

4.7. Modelado con redes neuronales LSTM

En lo que a los parámetros del modelo de red neuronal, cabe destacar que se han probado varias combinaciones de configuraciones, ya que un entrenamiento muy largo podía causar un sobreajuste de la red neuronal. Esto mismo lo podemos apreciar en las siguientes imágenes,

donde la red neuronal ha estado mejorando las predicciones, pero a la hora de realizar las predicciones futuras realiza una sobreestimacion.

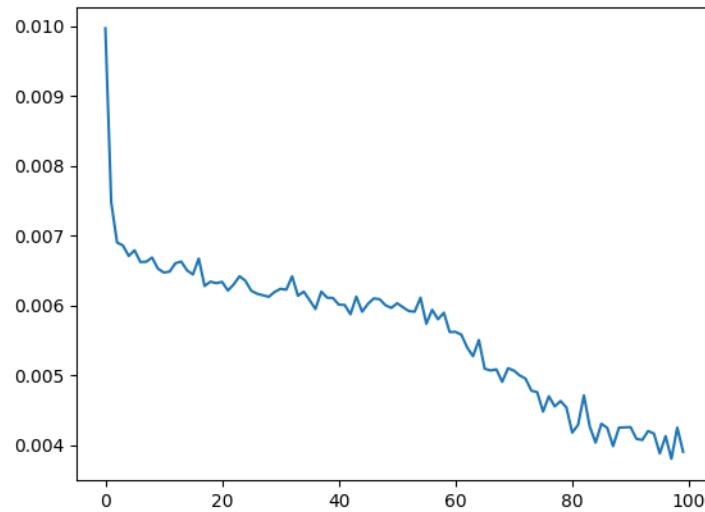


Figura 4.8: Evolución del modelo con un periodo de entrenamiento largo

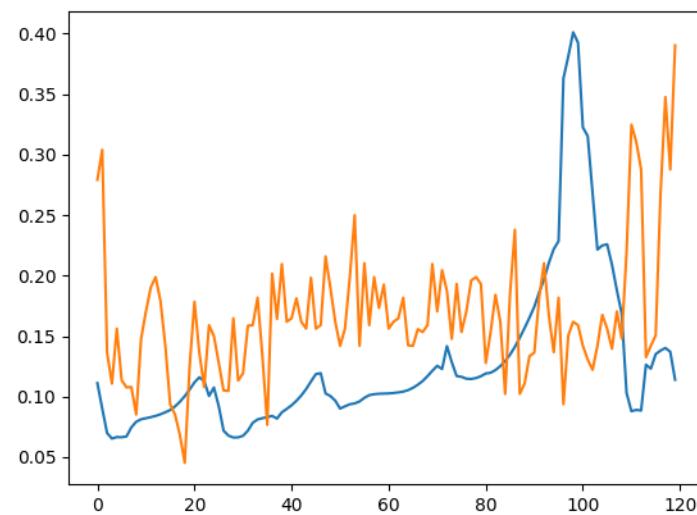


Figura 4.9: Predicciones del modelo LSTM con un periodo de entrenamiento largo

Por el contrario, al emplear un periodo de entrenamiento más corto se puede apreciar como las predicciones son más rígidas, pero se captura y se sigue de mejor manera la evolución de la contaminación.

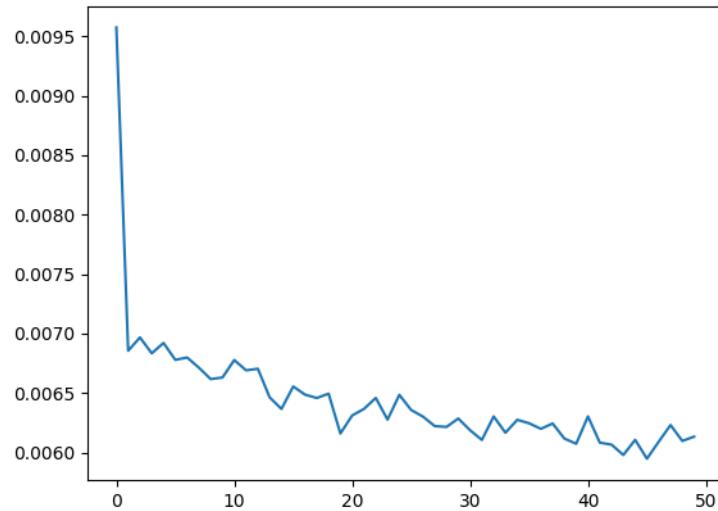


Figura 4.10: Evolución del modelo con un periodo de entrenamiento corto

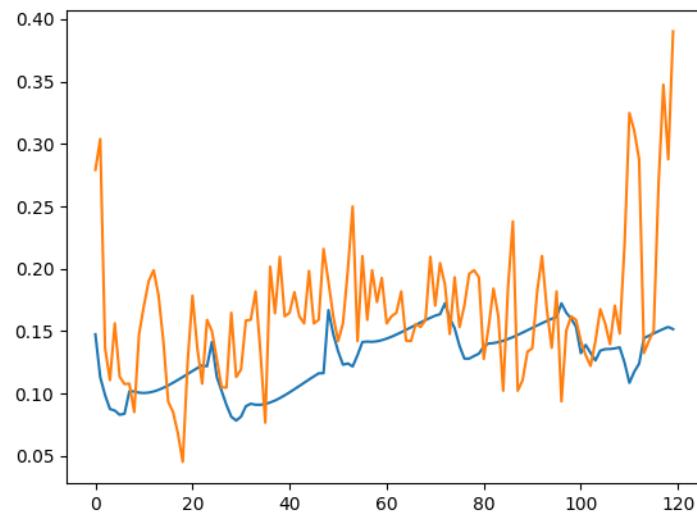


Figura 4.11: Predicciones del modelo LSTM con un periodo de entrenamiento corto

4.8. Comparación modelos

Por lo general, los tres modelos ajustan bien y son válidos, aunque a la hora de hacer las predicciones se puede apreciar como los modelos de series temporales clásicos se ven altamente influenciados por el comportamiento de los movimientos de personas, de forma que cuando no existe movimiento la contaminación decrece, pero esto no es cierto, ya que por su propia naturaleza, es muy difícil que desaparezca.

Por otra parte, el modelo de redes neuronales es mucho más flexible, pero se comporta de una forma muy lineal. En la siguiente imagen se pueden apreciar todas las predicciones así como los valores observados.

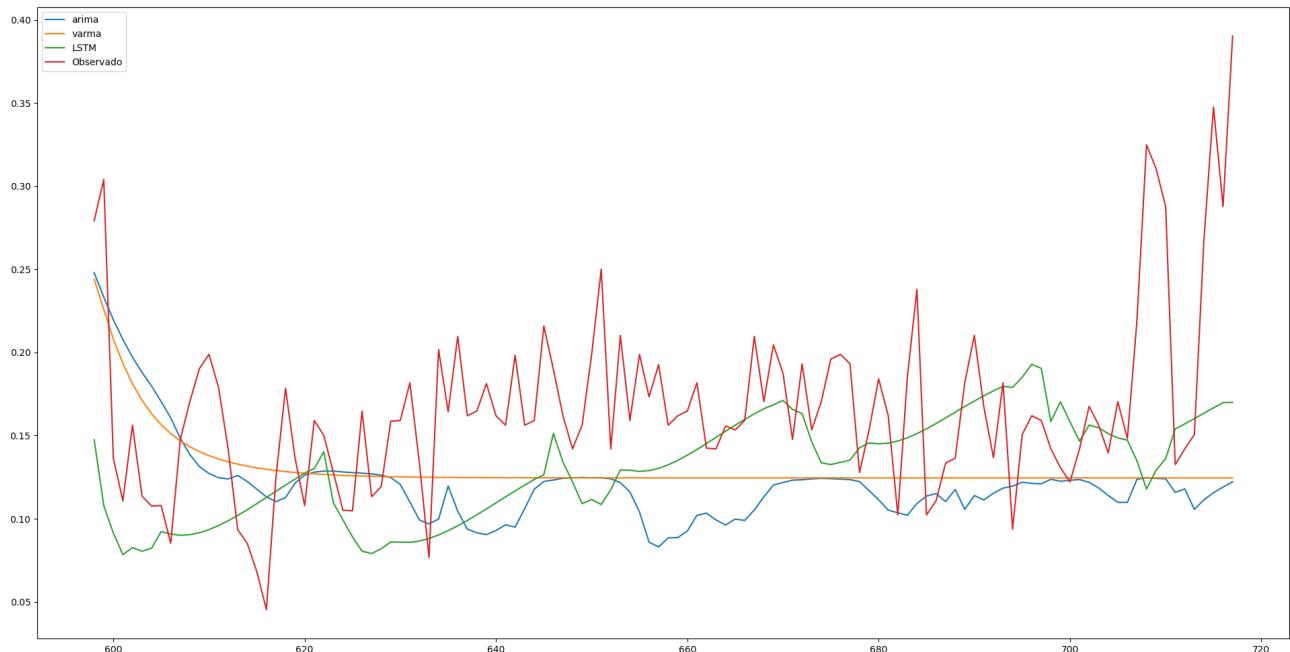


Figura 4.12: Predicciones de todos los modelos y los valores reales observador

Claramente salta a la vista como el modelo VARMA tiene unas predicciones lineales. Por otro lado los dos modelos restantes parecen que se podrían complementar para dar mejores resultados, es por ello, que si aplicamos un modelo de ensamblado sobre las propias predicciones de ambos modelos se podrían obtener mejores resultados.

4.9. Ensamble de predicciones

Para realizar el ensamble de las predicciones, se ha empleado únicamente las predicciones del modelo ARIMA y LSTM. Siendo los resultados de los diferentes ensambles probados:

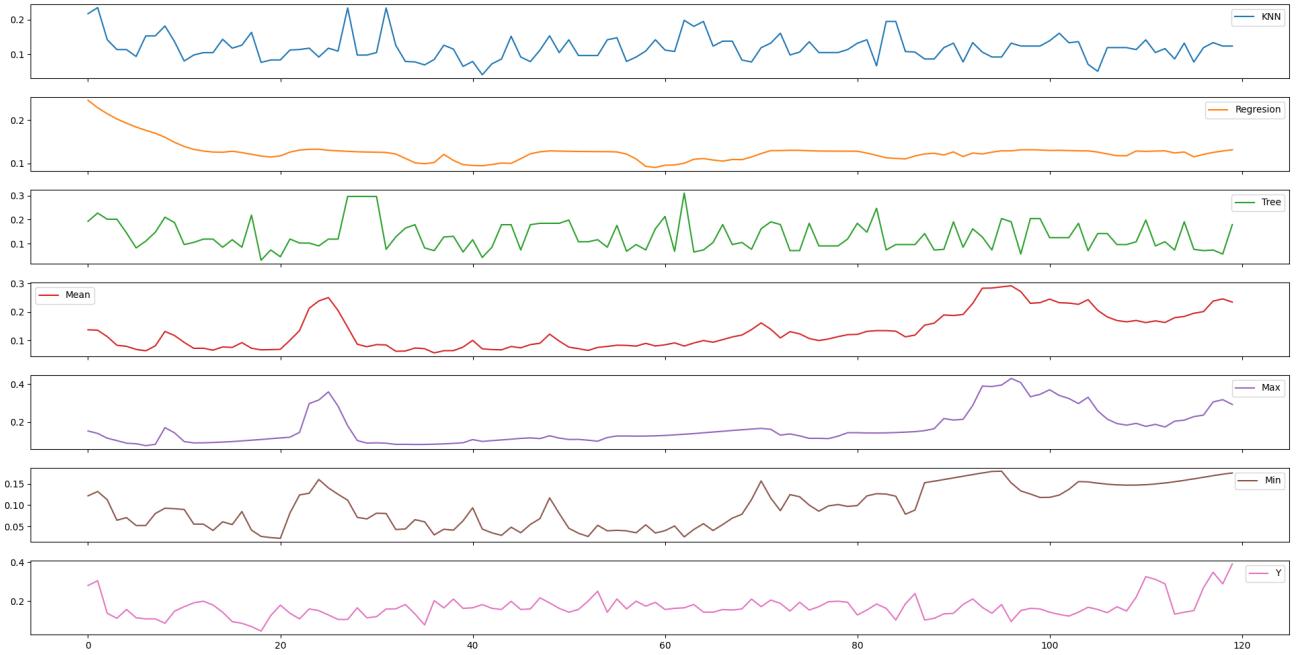


Figura 4.13: Ensamble de predicciones del modelo ARIMA y LSTM. De arriba a abajo: KNN, regresión lineal, árboles de decisión, media, máximo, mínimo y observado por la estación meteorológica

Una vez vistas las predicciones de los modelos ensamblados, pasaremos a ver las medidas de ajuste para todos los modelos ensamblados, siendo estos:

	BIC	MSE	SSE	MAE	AIC	AICC
KNN	-3980.73237	0.00627	0.75236	7.55453	4.56907	0.09215
Regresion	-4104.27014	0.00510	0.61194	6.68712	4.98224	0.10048
Tree	-3972.72281	0.00635	0.76251	7.78870	4.54228	0.09161
Mean	-3902.35625	0.00715	0.85772	8.76750	4.30694	0.08686
Max	-3718.87600	0.00971	1.16573	8.89530	3.69330	0.07449
Min	-3739.28515	0.00939	1.12662	9.47202	3.76156	0.07586

A la vista de la tabla anterior, observamos como el modelo ensamblado de regresión lineal presenta mejores resultados en todas las métricas, cosa que ya se sospechaba debido a que mezcla los ciclos y picos de la contaminación.

4.10. Selección del mejor modelo

Una vez ya tenemos creados los modelos originales y ensamblados, hablaremos sobre el mejor modelo de todos y si realmente es necesario aplicar un modelo de ensamblado. Para ello, se emplearan tablas descriptivas de los errores y errores absolutos así como sus distribuciones.

En primer lugar, hablaremos de los errores en general, los cuales vemos representados en los siguientes histogramas:

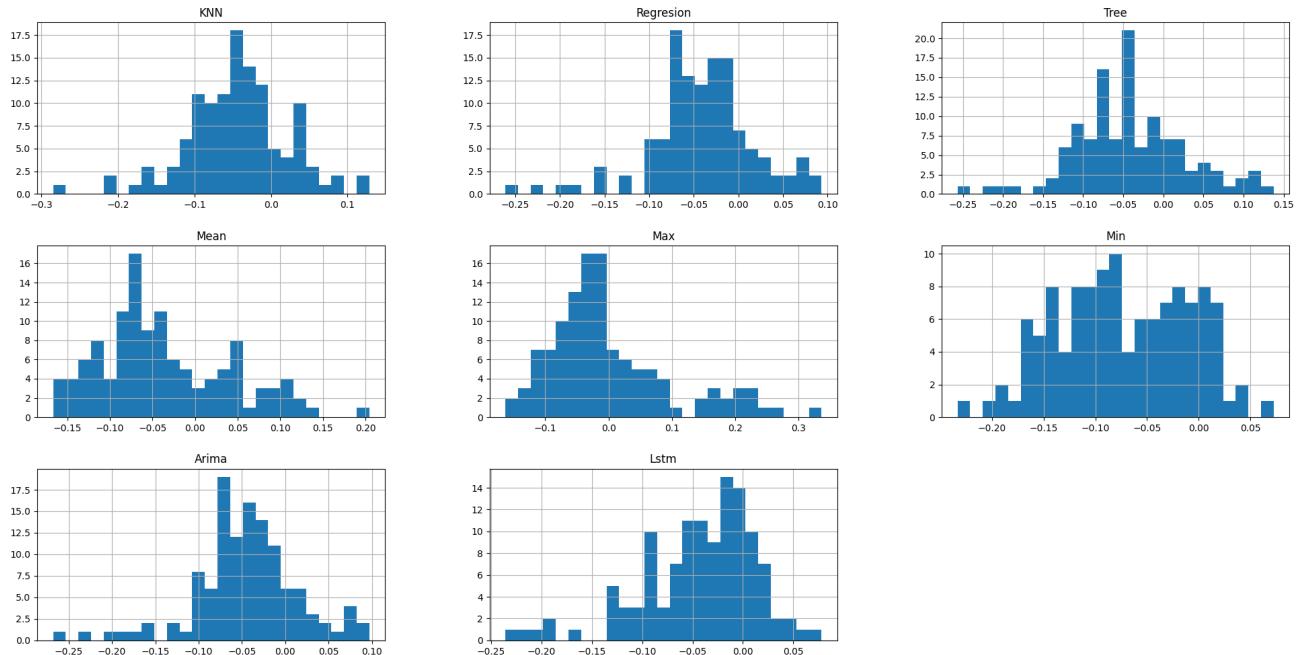


Figura 4.14: Residuos de los modelos originales y ensamblados

	KNN	Regresion	Tree	Mean	Max	Min	Arima	Lstm
count	120	120	120	120	120	120	120	120
mean	-0.0442	-0.0411	-0.0415	-0.0358	0.0017	-0.0733	-0.0454	-0.0432
std	0.0659	0.0586	0.0684	0.0769	0.0990	0.0636	0.0599	0.0573
min	-0.2849	-0.2613	-0.2567	-0.1667	-0.1630	-0.2331	-0.2680	-0.2360
25 %	-0.0798	-0.0689	-0.0822	-0.0843	-0.0607	-0.1205	-0.0723	-0.0707
50 %	-0.0427	-0.0380	-0.0444	-0.0511	-0.0216	-0.0807	-0.0429	-0.0338
75 %	-0.0124	-0.0124	0.0000	0.0145	0.0388	-0.0197	-0.0179	-0.0045
max	0.1286	0.0932	0.1377	0.2044	0.3365	0.0724	0.0970	0.0781

Si nos fijamos en los rangos de los errores, vemos como el modelo de regresión presenta unos errores concentrados y con un rango de errores pequeño comparado al resto de modelos, los cuales tienen distribuciones mas anchas y asimétricas.

Por otra parte, si nos fijamos en los errores absolutos, se puede apreciar a simple vista que el modelo ARIMA, LSTM y el ensamblado de la media concentran los errores a la izquierda, lo que indica que sus errores están más controlados. Además, los perceptibles son más bajos que el resto de modelos, lo que indica que los tres modelos mencionados tienen menos probabilidades de cometer errores más altos.

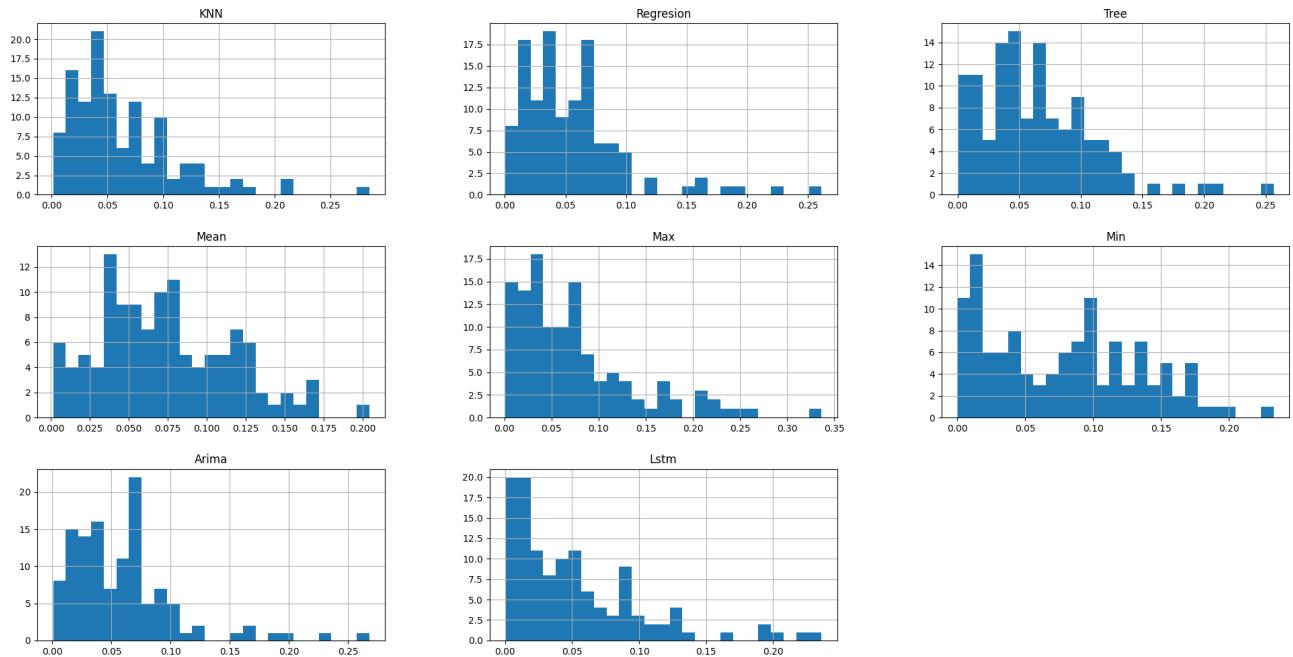


Figura 4.15: Residuos en valor absoluto de los modelos originales y ensamblados

	KNN	Regresión	Tree	Mean	Max	Min	Arima	Lstm
count	120	120	120	120	120	120	120	120
mean	0.0630	0.0557	0.0649	0.0731	0.0741	0.0789	0.0595	0.0518
std	0.0482	0.0448	0.0465	0.0427	0.0652	0.0564	0.0458	0.0497
min	0.0013	0.0003	0.0000	0.0014	0.0009	0.0001	0.0007	0.0002
25 %	0.0307	0.0242	0.0333	0.0408	0.0294	0.0262	0.0279	0.0158
50 %	0.0505	0.0488	0.0593	0.0685	0.0578	0.0807	0.0540	0.0387
75 %	0.0837	0.0709	0.0922	0.1043	0.0962	0.1205	0.0742	0.0761
max	0.2849	0.2613	0.2567	0.2044	0.3365	0.2331	0.2680	0.2360

A la vista de todo lo anterior el ensamblado de la regresión obtienen mejores resultados ya que aúna lo mejor de los modelos originales, es decir, los ciclos del modelo ARIMA y los picos del modelo LSTM.

Capítulo 5

Conclusiones

A la vista del trabajo queda patente que los IMSIS presentes en el área aportan cierta información sobre la contaminación ambiental y se podría reforzar con más factores tales como el viento, humedad, lluvia, etc... ya que dichos agentes externos son capaces de modificar la concentración de las partículas de suspensión en el aire. Es por ello que si alguien quisiera replicar este mismo trabajo, sería recomendable invertir más tiempo y esfuerzo en mejorar la estación meteorológica u optar por comprar una estación profesional (en el caso de que se opte por una vía close source). Por otra parte, también sería aconsejable monitorear más antenas de otros operadores, lo que implica comprar más SDR-RTL ya que maximizaría la captación de otros teléfonos así como poder monitorizar de forma más eficaz el tráfico rodado.

En lo que a los modelos respecta, sería necesario ampliar el horizonte de estudio, ya que únicamente se ha trabajado con 1 mes de datos y esto permitiría tener más muestras y compensar la pérdida o corrupción de los datos.

Una vez aclarado las limitaciones del trabajo y sus posibles subsanaciones, toca enfocarse en las conclusiones finales de este trabajo. Para ello, se entrenaran los modelos ganadores con los datos de la serie temporal de 25 micrómetros y se estudiaran los residuos con el fin de contrastar los resultados.

Como se puede observar en los siguientes gráficos y tablas de los residuos normales y absolutos, se aprecia un comportamiento similares a los obtenidos en los modelos aplicados a los datos de 100 micrómetros pero con una escala numérica diferentes.

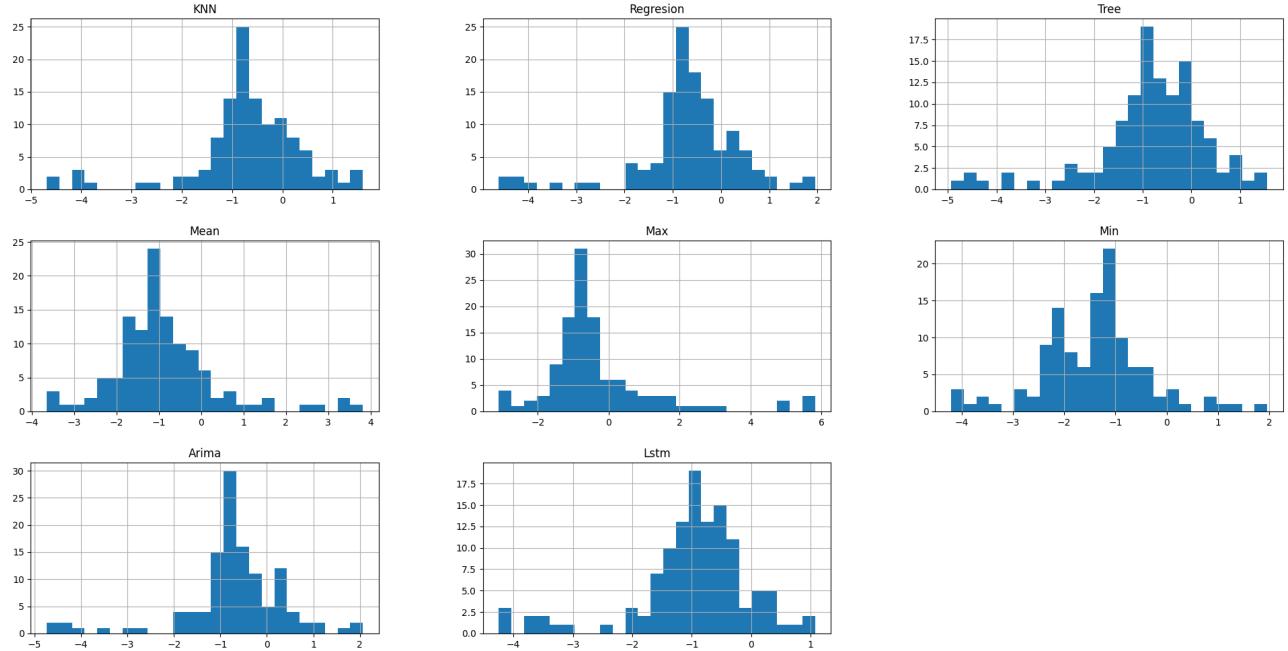


Figura 5.1: Residuos en valor absoluto de los modelos originales y ensamblados

	KNN	Regresion	Tree	Mean	Max	Min	Arima	Lstm
count	120	120	120	120	120	120	120	120
mean	-0.69020	-0.68564	-0.81813	-0.87843	-0.32692	-1.42993	-0.71950	-0.99092
std	1.11243	1.10551	1.17855	1.27674	1.62259	1.04419	1.13583	0.98961
min	-4.68731	-4.60490	-4.93310	-3.65220	-3.09924	-4.20517	-4.74057	-4.24477
25 %	-1.08800	-1.03693	-1.20809	-1.57486	-1.12651	-2.07738	-1.08148	-1.28823
50 %	-0.67674	-0.66562	-0.69788	-1.04421	-0.76352	-1.35485	-0.70519	-0.86315
75 %	-0.07926	-0.15851	-0.08453	-0.41658	-0.01948	-0.93626	-0.15013	-0.44882
max	1.59934	1.95075	1.55626	3.81925	5.82159	1.96427	2.06645	1.07234

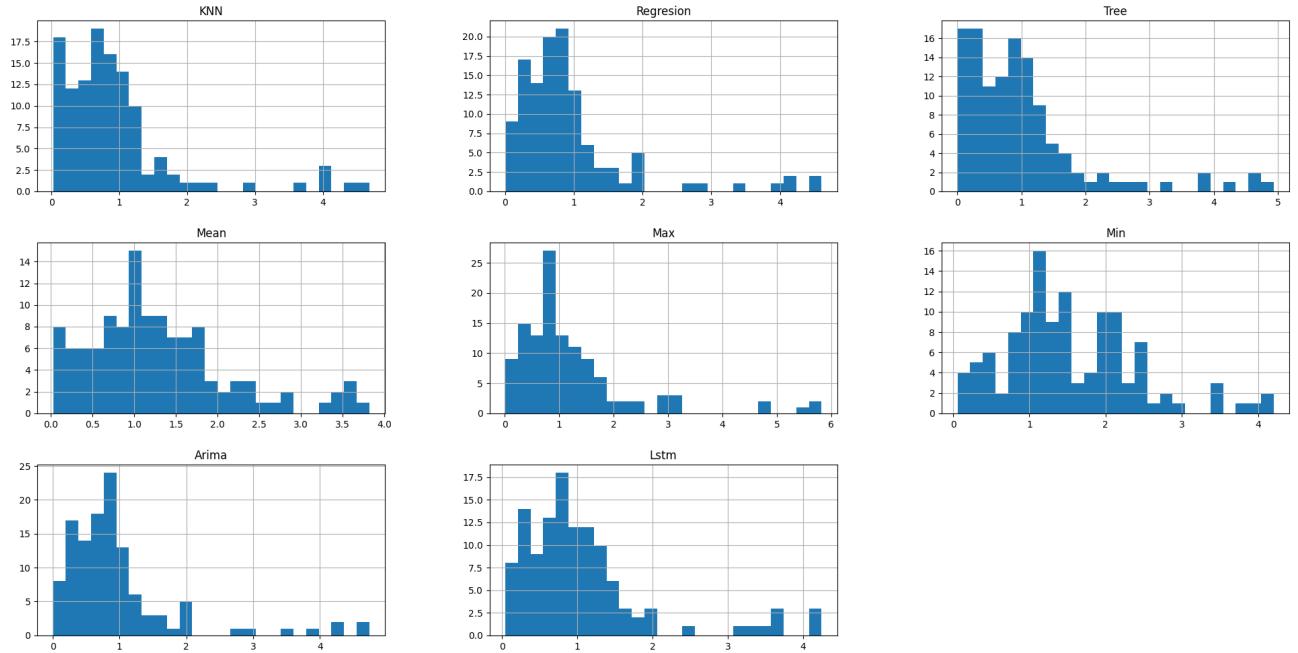


Figura 5.2: Residuos en valor absoluto de los modelos originales y ensamblados

	KNN	Regresion	Tree	Mean	Max	Min	Arima	Lstm
count	120	120	120	120	120	120	120	120
mean	0.94100	0.93883	1.02784	1.28789	1.22492	1.54015	0.98212	1.07862
std	0.90828	0.89854	0.99931	0.85767	1.10794	0.87192	0.91623	0.89235
min	0.02328	0.00278	0.00284	0.02518	0.01196	0.05705	0.00492	0.04335
25 %	0.40544	0.41176	0.37911	0.68251	0.56746	1.00085	0.43576	0.52596
50 %	0.74574	0.73589	0.82825	1.14889	0.89928	1.37249	0.78641	0.87935
75 %	1.11046	1.07594	1.28112	1.70078	1.43645	2.07738	1.10883	1.28823
max	4.68731	4.60490	4.93310	3.81925	5.82159	4.20517	4.74057	4.24477

En ambos casos se observa como el error del modelo es aproximadamente de una partícula, donde la media de partículas de 25 micrómetros en el ambiente es de 5 a 6 como se puede apreciar en la figura 4.2. Es por ello que el error que presenta los parámetros para el modelo final oscila entre el 16 % y 20 % de los datos reales.

En resumen, a la vista de los modelos de series temporales se ve como la concentración de teléfonos en la zona tiene un efecto a largo plazo. En el modelo ARIMA el valor asociado al parámetro de teléfonos tiene un signo negativo, mientras que en el modelo VARMA se

aprecia como como el retardo de una hora ya implica un signo positivo, lo que implica que a mayor concentración de teléfonos en la zona, mayor es la contaminación ambiental.

Bibliografía

- [1] antirez. Dump1090. <https://github.com/antirez/dump1090>.
- [2] George Box and Gwilym Jenkins. *Time series analysis; forecasting and control*. Holden-Day, San Francisco, 1970.
- [3] Pablo Saro Buendía. Isac-nav. *TFG (Facultad de Informatica)*, 2021.
- [4] Gonzalo José Carracedo Carballal. Sigdigger. <https://batchdrake.github.io/SigDigger/>.
- [5] François Chollet et al. Keras. <https://keras.io>, 2015.
- [6] Ministerio de Asuntos Económicos y Transformación Digital. Orden etd/1449/2021, de 16 de diciembre, por la que se aprueba el cuadro nacional de atribución de frecuencias. *Boletín Oficial del Estado*, 2021.
- [7] Grupo de Radio Hacking UCM. Oficina de software libre. <http://alapont.ucm.es/>.
- [8] Oficina de Software Libre UCM. Oficina de software libre. <https://www.ucm.es/oficina-de-software-libre>.
- [9] Fiscalía General del Estado. Circular 1/2013, de 11 de enero, sobre pautas en relación con la diligencia de intervención de las comunicaciones telefónicas. *Boletín Oficial del Estado*, 2013.
- [10] merbanan. Rtl 433. https://github.com/merbanan/rtl_433.
- [11] Olimex. Technical data. mq-135 gas sensor. *Olimex resources*, 2022.
- [12] Oros42. Imsi-catcher. <https://github.com/Oros42/IMSI-catcher>.

- [13] Ishtiaq Rouf, Rob Miller, Hossen Mustafa, Travis Taylor, Sangho Oh, Wenyuan Xu, Marco Gruteser, Wade Trappe, and Ivan Seskar. Security and privacy vulnerabilities of in-car wireless networks: a tire pressure monitoring system case study. *USENIX Association*, 2010.
- [14] Sección 1^a) Tribunal Supremo (Sala de lo Penal. Sentencia núm. 249/2008 de 20 mayo. recurso de casación núm. 10983/2007. *Boletin Oficial del Estado*, 2007.
- [15] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.
- [16] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Apéndice A

Bases legales del trabajo

Este trabajo hace uso de *IMSI catchers* pasivos y con fines meramente académicos y médicos. En lo que a consultas legales respecta, se consultaron diferentes sentencias publicadas en el Boletín Oficial del Estado.

En el documento de la Fiscalía General del Estado⁹ podemos apreciar los siguientes extractos donde se indica que la captura de los IMSIS no vulnera ningún derecho de la privacidad así como que es imposible saber la identidad de una persona únicamente con el IMSI, ya que sería necesario monitorizar las conexiones a gran escala y seguir físicamente al objetivo hasta triangular su posición y su IMSI con las antenas.

[...] Uno de estos derechos fundamentales es el derecho al secreto de las comunicaciones, proclamado ya por la Asamblea Nacional francesa en 1790: le secret des lettres est inviolable. Su reconocimiento tiene lugar al máximo nivel, en el art 18.3 CE, conforme al que "se garantiza el secreto de las comunicaciones y, en especial, de las postales, telegráficas y telefónicas, salvo resolución judicial". Desde una perspectiva internacional el derecho es reconocido en los arts. 12 de la Declaración Universal de Derechos Humanos de 10 de diciembre de 1948, 17 del Pacto Internacional de Derechos Civiles y Políticos de 19 de diciembre de 1966 y 8 del Convenio Europeo de Derechos Humanos y de las Libertades Fundamentales de 4 de noviembre de 1950. También más recientemente ha sido reconocido por el art 7 de la Carta de Derechos Fundamentales de la Unión Europea que dispone bajo la rúbrica Respeto de la vida privada y familiar que toda persona tiene derecho al respeto de su vida privada y familiar, de su domicilio y de sus comunicaciones. Estos textos constituyen parámetros para la interpretación de los derechos fundamentales y libertades (art. 10.2 CE). [...]

[...] Tanto el IMEI como el IMSI carecen de capacidad de información sobre la identidad del usuario, teniendo valor probatorio únicamente si se asocia a otros datos en poder de las operadoras [...]

[...] Ni el IMSI, ni el IMEI por sí solos, son datos integrables en el concepto de comunicación.[...]

Fiscalía General del Estado⁹

En otro extracto del Tribunal Supremo¹⁴, se indica que los IMSIS pueden considerarse dato personal siempre y cuando se relacionen con otros datos que permitan su deanonymización, pero debido a la naturaleza de este trabajo, no se consultan datos externos con el fin de deanonymizar dichos identificadores. Hasta el punto que al agregar los IMSIS en variables temporales se pierde la opción de ver IMSIS individuales.

[...] Admitido que esa numeración IMSI es integrable en el concepto de dato personal, por cuanto que mediante su tratamiento automatizado y su interrelación con otros datos en poder del operador puede llegar a obtenerse, entre otros datos, la identidad del comunicante, obligado resulta precisar el régimen jurídico de su cesión y, sobre todo, el de su aprehensión mediante acceso.[...]

Tribunal Supremo¹⁴