UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Predicción e interpolación dinámica de los niveles de contaminación atmosférica mediante datos de intensidad de tráfico y dirección del viento

TRABAJO FIN DE MÁSTER

Máster Universitario en Gestión de la Información

Autor: Lidia Contreras Ochando

Tutor: Cèsar Ferri Ramírez

Curso 2015-2016

"...Vivimos en el fondo de un mar de aire..."
Evangelista Torricelli (1608-1647)

Agradecimientos

Son muchas las personas que me han ayudado, apoyado o aconsejado durante estos dos años, tanto para el trabajo que se describe en esta memoria como para otros proyectos. Aunque estas pocas líneas no sean suficientes, mi agradecimiento es para todos ellos.

En primer lugar, mi agradecimiento va dirigido al Ayuntamiento de Valencia, en especial a Ramón Ferri y Ruth López, y a la empresa BSG Ingenieros, por proporcionarme el soporte y los datos con los que ha sido posible realizar este trabajo.

A los profesores y compañeros del Máster, por todo lo que he aprendido con ellos.

A Amparo López, del Departamento de Ingeniería Hidráulica y Medio Ambiente y a Mónica Catalá, del Departamento de Química, por explicarme cómo calcular los índices de la calidad del aire correctamente y ayudarme a comprender el comportamiento de los contaminantes en la atmósfera y su relación con el viento.

A Nicolas Lachiche, de la Universidad de Estrasburgo, por darme la oportunidad de continuar mi trabajo en otro país.

A Eduardo Vendrell, director de la ETSINF, por su confianza y recomendaciones, y por hacer que mi trabajo haya viajado hasta China.

A los profesores y compañeros del grupo ELP, del Departamento de Sistemas Informáticos y Computación, en especial a los miembros del grupo DMIP, José Hernández-Orallo y MªJosé Ramírez, por haberme acogido y ayudado durante el último año, y a Nando, por ejercer de hermano mayor científico.

A mis padres, que me animaron a volver a estudiar cuando todo parecía perdido y me han acompañado siempre que me he sentido sola, aunque estuvieran lejos.

A mi hermano Fran y a Cristina, que me acompañaron en el HackForGood donde nació está idea, porque sin ellos no habría sido posible.

Y sobre todo, a César Ferri, por todo lo que ha hecho por mi estos dos años; por compartir conmigo sus conocimientos, su experiencia, sus contactos y su tiempo; por dejarme utilizar la imaginación, hacerme dudar e incitarme a participar en todo; y por confiar en mí desde el principio de esta etapa y seguir confiando en mí para la siguiente.

¡Gracias a todos!

Lidia =)

Resumen

En este trabajo se presenta un método para predecir e interpolar los niveles de contaminación atmosférica en la ciudad de Valencia. En primer lugar, se comparan diferentes modelos de regresión, siendo capaces de predecir el nivel de cuatro contaminantes (NO, NO_2, O_3, SO_2) en las seis estaciones de medición de contaminación de la ciudad de Valencia. La fuerza y dirección del viento son factores clave en la propagación de los contaminantes, generados en gran medida por las emisiones producidas por los vehículos que circulan por las ciudades. Por esta razón, se estudian diferentes técnicas para incorporar estos factores en los modelos de predicción. En segundo lugar, se analiza como extrapolar las predicciones a toda la ciudad. Con este propósito, se propone un nuevo método de interpolación que tiene en cuenta la dirección del viento a la hora de calcular el resultado. Los experimentos con validación cruzada muestran que este método mejora los resultados en comparación con otros métodos conocidos. Finalmente, se utilizan estos métodos para extrapolar los resultados a toda la ciudad y generar mapas de la contaminación atmosférica en Valencia.

Palabras clave: Contaminación atmosférica, minería de datos, interpolación espacial

Resum

En este treball es presenta un mètode per a predir i interpolar els nivells de contaminació atmosfèrica en la ciutat de València. En primer lloc, es comparen diferents models de regressió, sent capaços de predir el nivell de quatre contaminants (NO, NO_2, O_3, SO_2) en les sis estacions de mesurament de contaminació de la ciutat de València. La força i direcció del vent són factors clau en la propagació dels contaminants, generats en gran manera per les emissions produïdes pels vehicles que circulen per les ciutats. Per esta raó, s'estudien diferents tècniques per a incorporar estos factors en els models de predicció. En segon lloc, s'analitza com extrapolar les prediccions a tota la ciutat. Amb este propòsit, es proposa un nou mètode d'interpolació que té en compte la direcció del vent a l'hora de calcular el resultat. Els experiments amb validació encreuada mostren que este mètode millora els resultats en comparació amb altres mètodes coneguts. Finalment, s'utilitzen estos mètodes per a extrapolar els resultats a tota la ciutat i generar mapes de la contaminació atmosfèrica a València.

Paraules clau: Contaminació atmosfèrica, minería de dades, interpolació espacial

Abstract

This work presents a method for predict and interpolate the levels of urban air pollution for the city of Valencia. First, we compare several regression models able to predict the levels of four different pollutants (NO, NO_2, O_3, SO_2) in the six pollution measurement stations of the city of Valencia. Wind Strength and Wind Direction are key features in the propagation of pollutants, generated mostly by vehicles circulating in the city. We study different techniques to incorporate these factors in the regression models. In second place, we analyse how to interpolate forecasts all around the city. Here, we propose a new interpolation method that takes wind direction into account. We compare this proposal with respect to well-known interpolation methods. By using these contamination estimates, we are able to generate pollution maps of the city of Valencia.

Key words: Air pollution, data mining, spatial interpolation

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

Existe una evidencia inequívoca de que las actividades humanas afectan a las condiciones climáticas del planeta y a la salud humana [13], sobre todo a partir de la gran concentración de gases de efecto invernadero en la atmósfera que causa el aumento en la temperatura media mundial. En el último cuarto de siglo el 60 % de los principales ecosistemas del mundo se ha degradado o utilizado de modo insostenible, según el Programa de las Naciones Unidas para el Medio Ambiente (PNUMA). Las ciudades acaparan actualmente el 75 % del consumo energético y son responsables del 75 % de las emisiones de carbono. La evaluación, control y mitigación de estas emisiones suponen un gran desafío, pero son fundamentales para revertir este proceso y procurar un mejor bienestar [38]. Según datos de la Organización Mundial de la Salud (OMS), mediante la disminución de los niveles de contaminación del aire se puede reducir la mortandad derivada de accidentes cerebrovasculares, cánceres de pulmón y neumopatías crónicas y agudas, entre ellas el asma [6].

La contaminación atmosférica es el resultado de la emisión de gases y partículas procedentes de un amplio conjunto de actividades tanto naturales como antropogénicas, es decir, actividades humanas. Este tipo de contaminación es uno de los factores con mayor impacto en la salud de las personas y animales, así como en el deterioro de ecosistemas vegetales y acuáticos [60]. La exposición prolongada a ambientes contaminados afecta a la calidad y condiciones básicas de vida e incrementa el riesgo de sufrir trastornos respiratorios, como neumonía, o enfermedades crónicas, como cáncer, insuficiencias cardiovasculares [92] o alzheimer [1][75]. Un trabajo reciente [90] relaciona cambios estructurales en el cerebro con las largas exposiciones a estos tipos de ambientes. El informe SOER 2015 [88] concluye que, a pesar de que la atmósfera en Europa ha mejorado en las últimas décadas, aún quedan muchas trazas de los contaminantes más peligrosos. Este informe estima que aproximadamente 430.000 ciudadanos murieron en 2011 a causa de la polución. Solo en España, la exposición prolongada a partículas finas en suspensión causa más de 25.000 muertes prematuras cada año [39]. En este contexto, se debe tratar de reducir la exposición al aire contaminado tanto como sea posible. Esto es especialmente importante para la población en riesgo, como niños, gente mayor, personas asmáticas o mujeres embarazadas.

El uso, tanto de estaciones de medición de polución, como de modelos de difusión atmosférica para la predicción de los niveles de contaminantes en zonas sin estaciones, es algo habitual, incluso recogido en el BOE como herramientas esenciales para conocer los niveles de contaminación. De hecho, en el R.D. 102/2011, de 28 de enero, relativo a

la mejora de la calidad del aire[1], donde se recogen los valores límite para los diferentes contaminantes, aparecen también unos umbrales superior e inferior de evaluación para cada contaminante. Estos umbrales establecen la necesidad o no de emplear mediciones fijas o técnicas de modelización o combinaciones de ambos.

Estos modelos de difusión atmosférica son modelos predictivos a los que se le suministran unos datos de entrada (datos del emisor, del receptor y meteorológicos) a partir de los cuales genera unos datos de salida (concentraciones horarias, diarias, etc., de los distintos contaminantes) que varían según la complejidad del modelo empleado y constituyen el eslabón de enlace entre la fuente y el receptor, simulando los procesos físicos y químicos que sufren los contaminantes durante su transporte. Una representación correcta de estos fenómenos requiere la solución de diferentes sistemas de ecuaciones diferenciales en derivadas parciales, lo cual constituye un problema matemático de dimensiones considerables, que requiere grandes capacidades de computo [60].

Por otra parte, el volumen y variedad de información que se encuentra informatizada en bases de datos y la inmensa cantidad de datos producidos por los sensores, ha crecido espectacularmente en las últimas décadas. Mucha de esta información almacena históricos sobre diferentes mediciones, lo cual resulta útil para explicar el pasado, entender el presente y predecir información futura. Además, los datos pueden provenir de diversas fuentes, por lo que es necesario un análisis de los mismos para la obtención de información útil [42].

Estos problemas y limitaciones han motivado la creación de nuevas herramientas y técnicas que permitan extraer conocimiento útil desde la información disponible. Todas estas herramientas y técnicas se engloban dentro del concepto de Minería de Datos, cuyo resultado son conjuntos de reglas, ecuaciones, arboles de decisión..., que permiten construir modelos predictivos que ayudan en el cálculo y toma de decisiones en muy variados contextos y con diferentes aplicaciones como, por ejemplo, la estimación de cuánta contaminación se producirá dentro de una hora.

El ejemplo más avanzado de esta combinación de técnicas en España, enfocadas al análisis de la calidad del aire, podemos verlo en el Sistema CALIOPE[2]. Este sistema, creado por el Departamento de Ciencias de la Tierra del Barcelona - Centro Nacional de Supercomputación (BSC-CNS), realiza pronósticos de los niveles de polución para España, utilizando para ello datos de estaciones de medición de contaminación de todo el país. CALIOPE realiza sus predicciones mediante regresión lineal y extrapola las estimaciones a zonas sin estaciones mediante el método de interpolación espacial *kriging*, ambos explicados en esta memoria.

El coste de realizar el análisis, desarrollo e implantación de un sistema de estas características requiere un gran despliegue de medios, tanto económicos como computacionales. Sin embargo, es posible realizar, con muchos menos recursos, un modelo a menor escala para, por ejemplo, una sola ciudad. Comenzar un sistema de estas características desde cero tiene ciertas ventajas, como el hecho de que permite probar los modelos implementados en otros sistemas y compararlos con otros modelos nuevos o poco usados. En este trabajo, abordaremos el problema de crear un sistema de predicción de conta-

[1] Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire: https://www.boe.es/diario_boe/txt.php?id=BOE-A-2011-1645

[2] Sistema CALIOPE: http://www.bsc.es/caliope/

minación probando diferentes técnicas, pero centrándonos únicamente en la ciudad de Valencia como caso de estudio, ciudad que por otra parte, no tiene un sistema público de predicción de contaminación.

El primer estudio sobre niveles de contaminación centrado en la ciudad de Valencia se realizó en 1986 y se llevó a cabo en la Universitat Politècnica de Valencia, bajo petición del Ayuntamiento de Valencia. Se trataba de un "modelo matemático urbano multifuente de dispersión atmosférica de contaminantes basado en la formulación gaussiana", mediante el cual se calcularon las concentraciones medias para el año 1986 en 100 puntos de estudio en el casco urbano de la ciudad. Para ello se sirvieron de datos de aforos de tráfico, fuentes de emisión de contaminantes, datos de contaminación de 10 años y datos meteorológicos, entre otros. Este estudio, que requirió algo más de dos años, concluyó que en la ciudad de Valencia el tráfico era responsable del 51.44 % de la contaminación por SO_2, del 87.37 % de contaminación por partículas en suspensión, del 87.64 % de la contaminación por NO_x y de toda la contaminación por plomo [60].

Según un informe de niveles de contaminación por país y ciudad, publicado en 2014 [10], Valencia se encuentra entre las 10 ciudades de España con los niveles más altos de contaminación, concretamente en los niveles de partículas en suspensión ($PM_{2,5}$). La comarca valenciana de l'Horta, donde se encuentra la ciudad de Valencia, superó en 2014 el valor límite anual de dióxido de nitrógeno (NO_2), según la última evaluación de la calidad del aire en España del Ministerio de Medio Ambiente [65]. Contaminante que es causado en su mayor parte por los motores de combustión de los automóviles, fundamentalmente los diésel.

En la ciudad de Valencia existen seis estaciones de medición de los niveles de contaminación atmosférica. Sin embargo, los datos que recogen estos sensores requieren una validación previa a su publicación que puede llegar a durar varias horas (3 horas en el caso de Valencia). Este retraso en la publicación de los datos puede representar un problema, ya que los niveles altos de contaminación no son detectados en tiempo real o con antelación, de modo que no pueden ser prevenidos para poder tomar medidas con antelación. Además de esto, la red de sensores tiende a ser limitada debido al alto coste de los equipos y su mantenimiento, haciendo que los resultados se concentren en unas pocas zonas de las ciudades.

La contaminación atmosférica es, por tanto, un tema que no debe ser subestimado. Es importante controlar de manera eficiente los niveles de contaminación atmosférica hasta su eliminación o hasta conseguir la reducción a niveles aceptables de aquellos agentes (gases, partículas en suspensión, elementos físicos y hasta cierto punto agentes biológicos) cuya presencia en la atmósfera puede ocasionar efectos adversos en la salud de las personas o en su bienestar, efectos perjudiciales sobre la vida de las plantas y de los animales, daños a materiales de valor económico para la sociedad y daños al medio ambiente. A no ser que se lleve a cabo un control adecuado, la multiplicación de las fuentes contaminantes del mundo moderno puede llegar a producir daños irreparables para el medio ambiente y para toda la humanidad [86].

1.1 Objetivos

Considerando las restricciones mencionadas, en este trabajo abordaremos el problema de producir predicciones de los niveles de contaminación para toda una ciudad, utilizando Valencia como caso de estudio. Con este propósito, estudiaremos cómo realizar la predicción de contaminación atmosférica, empleando para ello datos históricos de cuatro contaminantes (NO, NO_2, O_3, SO_2) obtenidos de las seis estaciones de medición de contaminación con las que cuenta Valencia y que proporciona la web de datos abiertos del Ayuntamiento, así como la web de la Generalitat Valenciana. Para generar estas predicciones, estudiaremos el resultado obtenido a partir de diferentes técnicas para construir modelos de regresión, entrenados mediante características que puedan originar y/o afectar a los niveles de contaminación, como la intensidad del tráfico o la observación meteorológica, de modo que podamos mejorar la precisión.

A continuación, plantearemos como interpolar estas predicciones para, de esta manera, ser capaces de mostrar la concentración aproximada de contaminantes en toda la ciudad. Con este fin, analizaremos métodos conocidos de interpolación espacial [58], como *Inverse Distance Weigthing* (IDW) o *Kriging*. Estos métodos son estáticos, es decir, no consideran las condiciones del contexto en los puntos a interpolar. Los parámetros meteorológicos (en especial el viento), pueden afectar potencialmente la manera en que los contaminantes se dispersan por la ciudad. Para corregir esto, propondremos un nuevo método de interpolación que utiliza la información del viento, con la intención de mejorar la interpolación, considerando esta aproximación como un método dinámico, sensible al contexto del punto a interpolar.

1.2 Estructura del documento

Esta memoria está estructurada como sigue:

- En el capítulo 2 se presentan los aspectos más significativos sobre la contaminación que se usarán a lo largo de este trabajo.
- El capítulo 3 describe el proceso de minería de datos y las técnicas de regresión, que se utilizaran en este proyecto.
- El capítulo 4 sirve para presentar la interpolación espacial y sus aplicaciones.
- En el capítulo 5 se describe la selección, limpieza y preparación de los datos que se usarán en los experimentos de este trabajo.
- El proceso de predicción de los contaminantes y los experimentos realizados con técnicas de regresión se describen en el capítulo 6.
- En el capítulo 7 se describen los experimentos realizados con el fin de extrapolar los contaminantes a toda la ciudad.
- En el capítulo 8 se presenta airVLC, una aplicación para mostrar tanto las predicciones como la extrapolación en un mapa interactivo de Valencia.
- El capítulo 9 contiene las conclusiones de este trabajo y el trabajo futuro.

CAPÍTULO 2
Contaminación atmosférica

En esta sección profundizaremos en ciertos aspectos relacionados con la contaminación atmosférica que se utilizan a lo largo de este trabajo, como pueden ser el efecto de la contaminación en la salud, la relación de los parámetros meteorológicos en la dispersión o concentración de los contaminantes o el marco legal que existe actualmente y que regula los niveles de contaminación límite.

2.1 Introducción

La atmósfera que rodea la Tierra está constituida por una mezcla de gases, entre los cuales destacan el nitrógeno, el oxígeno, el argón y el dióxido de carbono. Sin embargo, en cualquier tipo de atmósfera se puede detectar la presencia compuestos que la impurifican. Estos compuestos se consideran contaminantes cuando su concentración es excesivamente alta o son capaces de provocar un perjuicio notable en la vida sobre la tierra [60].

2.2 Tipos de contaminantes

En función de su origen los contaminantes pueden clasificarse en primarios y secundarios:

- Contaminantes primarios: Aquellas sustancias que son vertidas directamente a la atmósfera por fuentes emisoras, por ejemplo, monóxido de carbono (CO), óxidos de nitrógeno (NO_x) y óxidos de azufre (SO_x).

- Contaminantes secundarios: Son los que se producen como consecuencia de las transformaciones, a través de reacciones físicas y químicas, que sufren los contaminantes primarios en la atmósfera, por ejemplo, el ozono (O_3) [48].

2.3 Fuentes de emisión

Las fuentes de emisión de contaminación son muy numerosas, y pueden ser originados de forma natural o antropogénica.

Se denominan fuentes naturales a los procesos propios de la naturaleza a partir de los cuales se desprenden gases a la atmósfera, tales como erupciones volcánicas, incendios

naturales, etc. Por ejemplo, los granos de arena o las tormentas de polvo son contaminantes naturales por si mismos cuando el viento los arrastra y se convierten en partículas en suspensión.

Las fuentes naturales realizan una mayor aportación de gases en el seno de la atmósfera, sin embargo, son las emisiones originadas en áreas urbanas e industriales las más preocupantes, por producirse en núcleos donde hay mayor densidad de población y por concentrarse las emisiones en superficies poco extensas. Las fuentes antropogénicas son debidas a la actividad humana, originándose principalmente en los procesos de combustión de combustibles fósiles, procesos industriales, tratamientos y eliminación de residuos, etc.

2.3.1. Transporte terrestre

Las emisiones del transporte terrestre provienen de fuentes, causadas por el tráfico de vehículos por las carreteras. La quema de combustibles fósiles por los vehículos produce, sobre todo, monóxido de carbono (CO), Óxidos de nitrógeno (NO_x), dióxido de azufre (SO_2) y partículas en suspensión (PM). Según el origen de estas emisiones se pueden clasificar en [9]:

- Combustión, por el uso de combustible.
- Desgaste de neumáticos, frenos y pavimento.
- Resuspensión, por la acción de las ruedas sobre la superficie produciendo partículas en suspensión.

2.4 Ciclos

La variabilidad en los niveles de contaminación está influenciada, sobre todo, por el ciclo diurno.

En primer lugar, las emisiones, independientemente de la fuente, son menores por la noche debido a que las industrias y negocios están cerrados o reducen su actividad y en segundo lugar, hay un patrón diurno de transporte y emisión de contaminantes.

Por otro lado, se diferencian dos patrones: uno entre semana y otro en fin de semana, asociados con el cambio de vida de los fines de semana en comparación con el resto de la semana. Por último, hay un ciclo estacional, asociado con la diferencia en el clima y el tiempo durante las cuatro estaciones: primavera, verano, otoño e invierno. Los cambios climáticos afectan a la intensidad de la fuente, el transporte y la dispersión [18].

2.5 Transporte y dispersión

Los contaminantes del aire pueden ser transportados e incluso transformados en la atmósfera. La ubicación de los receptores con respecto a las fuentes y las influencias atmosféricas afectan a las concentraciones de los contaminantes y la sensibilidad de los receptores a estas concentraciones determina los niveles medidos.

Si el movimiento del aire a través de una fuente contaminante es lento, las concentraciones de contaminantes que se muevan a favor del viento serán mucho más altas que si el aire los moviera rápidamente, lejos de la fuente de emisión [18].

Los factores meteorológicos constituyen simples indicadores de la capacidad de dispersión de la atmósfera, pero ninguno de ellos por separado puede definir adecuadamente tal capacidad. Los factores más importantes a pequeña escala son la dirección y velocidad del viento y la turbulencia.

Estos factores determinan el grado de influencia que el transporte atmosférico tiene sobre la dispersión de los contaminantes. Otros factores meteorológicos son la radiación solar y la temperatura, que afectaran a las velocidades y dispersión de los contaminantes; la nubosidad, que altera la radiación solar y la precipitación, que elimina aerosoles y partículas a la vez que arrastra contaminantes y cuyo efecto suele prolongarse durante varios días [60].

2.5.1. Dirección del viento

La dirección inicial de transporte de los contaminantes desde la fuente viene determinada por la dirección del viento. Si el viento sopla directamente hacia un receptor (estación de medición), un cambio en la dirección de tan solo 5º puede hacer que las concentraciones medidas por el receptor desciendan entre un 10 % y un 90 %. La dirección de transporte de los contaminantes es muy importante en la evaluación del origen donde existen receptores sensibles a dos o más fuentes o al tratar de evaluar un modelo estadístico comparado con valores reales [18].

2.5.2. Velocidad del viento

Uno de los efectos de la velocidad del viento es diluir los contaminantes desde el punto de emisión, condicionando la rapidez con la cual el contaminante se separa de le fuente que lo ha originado. Si una fuente está en una superficie elevada, la dispersión se realizará en la dirección del viento (Ver Figura 2.1).

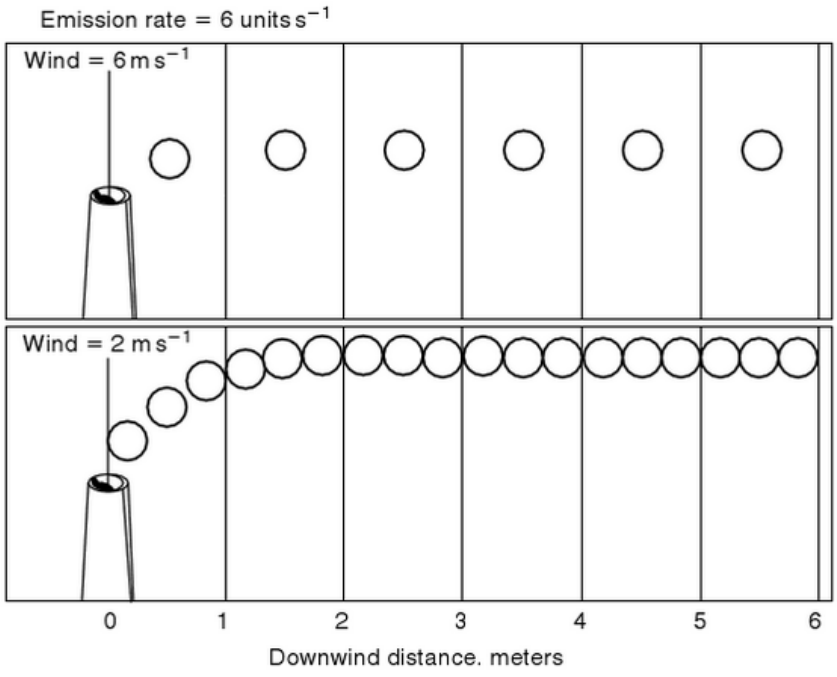

Figura 2.1: Dispersión de los contaminantes según la velocidad del viento (Fuente [18])

2.6 Efectos de la contaminación atmosférica

Los efectos de la contaminación atmosférica pueden ser considerados a dos niveles:

- Efectos directos de los contaminantes: La contaminación atmosférica a nivel local es fuente de numerosos problemas como la aparición de riesgos para la salud, asociados sobre todo a la inhalación de gases, deterioro de edificios y daños en la vegetación.
- Efectos indirectos de los contaminantes: La acumulación de gases en la atmósfera puede desestabilizar el equilibrio de la misma produciendo fenómenos a escala mundial como el calentamiento global, la acidificación de los suelos o el aumento de la radiación [48].

2.6.1. Efectos según la estación del año

Los efectos de la contaminación atmosférica sobre la salud pueden diferenciarse según la estación del año en la que nos encontremos (verano o invierno), ya que los cambios en las temperaturas alteran los contaminantes presentes en la atmósfera [13].

Los componentes principales en invierno son el SO_2 y las partículas en suspensión. El SO_2 es un gas irritante que puede producir broncoconstricción[1]. El resultado del estudio del proyecto APHEA [54] indica un incremento del 4 % en la mortalidad por incremento de 100 $\mu g/m^3$ en los niveles medios diarios de SO_2. Por su parte, el poder tóxico de las partículas en los ambientes urbanos procede de su alta composición en partículas ultrafinas, que pueden provocar un aumento de los problemas respiratorios e incrementar la coagulabilidad plasmática. Se asocian los niveles altos de partículas con la disminución del funcionamiento pulmonar.

La contaminación en verano procede principalmente de las reacciones de hidrocarburos y óxidos de nitrógeno, estimulados por la luz solar. El ozono es el componente más tóxico de esta estación, sin embargo, otros como los nitratos orgánicos y los aldehídos pueden provocar irritación ocular. La OMS clasifica los efectos esperados por exposiciones a concentraciones de ozono de 400 $\mu g/m^3$ y superiores como severas. Las personas que previamente padecen enfermedades respiratorias o deficiencias en el aparato circulatorio son más sensibles a trastornos de este tipo. Por estas razones, la Generalitat Valenciana dispone de un portal de aviso y recomendaciones diarias para altos niveles de Ozono en la Comunidad Valenciana[2].

2.7 Sistemas de vigilancia de la calidad del aire

Los sistemas de vigilancia están compuestos por redes de una o varias estaciones manuales o automáticas, donde se recogen muestras de los distintos contaminantes, que son analizadas para dar valores medios de concentración.

Según el ámbito geográfico de cobertura de las redes de control de la contaminación atmosférica, pueden clasificarse de la siguiente manera:

- Fondo urbano: Sirven para conocer los niveles de contaminación de una zona urbana. Estas estaciones no deben estar directamente influenciadas por fuentes locales, como tráfico o industrias. Representan niveles de exposición de los ciudadanos.

[1] La broncoconstricción es el estrechamiento de las vías respiratorias.
[2] Informe diario de Ozono GVA: http://www.habitatge.gva.es/web/informes-previozono

- Urbana de tráfico: Su finalidad es medir los niveles de contaminación en una vía con tráfico considerable. Están influenciadas directamente por las emisiones del tráfico y se deben ubicar dentro de una ciudad o en las cercanías de una carretera o autopista.
- Industrial: Se ubican cerca de una implantación industrial para controlar la contaminación producida por la misma.
- Fondo regional: Se ubican lejos de zonas urbanas o suburbanas. Su finalidad es determinar la contaminación atmosférica de fondo a nivel regional.

2.8 Marco legal de la calidad del aire

La normativa sobre calidad del aire [66] tiene su origen en la Directiva 96/62/CE del Consejo, de 27 de septiembre de 1996, sobre evaluación y gestión de la calidad del Aire Ambiente, o Directiva Marco, que plantea criterios, objetivos y técnicas de evaluación. Posteriormente, la mayor parte de estas normas fueron sustituidas por directivas hijas para los diferentes grupos de contaminantes.

2.8.1. Normativa Europea

La normativa europea sobre calidad del aire viene representada por las siguientes Directivas:

- Directiva 2008/50/CE del Parlamento Europeo y del Consejo, de 21 de mayo de 2008, relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa.

  Establece medidas para evitar, prevenir o reducir los efectos nocivos para la salud humana y el medio ambiente; evaluar la calidad del aire; obtener información para controlar la evolución a largo plazo; asegurar que la información está a disposición de los ciudadanos; mantener la calidad del aire en buenas condiciones y mejorarla y fomentar la cooperación entre los Estados miembros.

  Ha sido transpuesta en España mediante el Real Decreto 102/2001, de 28 de enero, relativo a la mejora de la calidad del aire.

- Directiva 2004/107/CE del Parlamento Europeo y del Consejo, de 15 de diciembre de 2004, relativa al arsénico, el cadmio, el mercurio, el níquel y los hidrocarburos aromáticos policíclicos en el aire ambiente.

  Establece un valor objetivo para las concentraciones de arsénico, cadmio, níquel y benzo(a)pireno; garantiza el mantenimiento de la calidad del aire; establece métodos y criterios comunes de evaluación y garantiza la obtención y puesta a disposición publica de información.

2.8.2. Normativa española

La legislación española sobre calidad del aire viene representada por las siguientes normas:

- Ley 34/2007, de 15 de noviembre, de calidad del aire y protección de la atmósfera.

  Tiene como fin alcanzar unos niveles óptimos de calidad del aire para evitar, prevenir o reducir riesgos o efectos negativos sobre la salud humana, el medio ambiente y demás bienes de cualquier naturaleza.

- Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire.

  Define y establece objetivos de calidad del aire con respecto a las concentraciones de dióxido de azufre, dióxido de nitrógeno y óxidos de nitrógeno, partículas, plomo, monóxido de carbono, ozono, arsénico, cadmio, níquel y benzo(a)pireno.

2.9 Índice de la calidad del aire

Los niveles de emisión e inmisión de la calidad del aire se hallan regulados mediante la especificación de límites máximos. La concentración máxima de emisión representa la mayor cantidad de contaminante que una fuente está autorizada a emitir. En el caso de la inmisión, el límite se denomina concentración máxima de inmisión y representa una limitación de tipo sanitario que regula la calidad del aire en cuestión. Suele estar acompañado en las normas legales por el máximo periodo de tiempo en que es autorizada su existencia [60].

Para indicar la calidad del aire en cada estación se emplean dos índices: Índice de Calidad Horario (ICH) e Índice de Calidad Diario (ICD). Se calculan a partir de los datos de las estaciones automáticas teniendo en cuenta 5 contaminantes: SO_2, NO_2, PM_{10}, CO y O_3. En la siguiente tabla se pueden ver las concentraciones asociadas al valor 100 de cada índice.

Directiva	*Contaminante*	*Concentración asociada al valor 100 de ICH* ($\mu g/m^3$)	*Concentración asociada al valor 100 de ICD* ($\mu g/m^3$)
Directiva Europea 1999/30/CE	SO_2	350	125
Directiva Europea 2000/69/CE	CO	10000	10000
Directiva Europea 1999/30/CE	NO_2	200	100
Directiva Europea 1999/30/CE	PM_{10}	75	50
Directiva Europea 2002/3/CE	O_3	180	120

Tabla 2.1: Concentraciones asociadas al valor 100 de los índices de calidad del aire ICH (índide de calidad horario) e ICD (índice de calidad diario).

Como se puede ver en la tabla 2.2, cualitativamente se establecen 4 tramos:

Valor IC	*Calidad*	*Contaminación*	*Color asociado*
0-50	Excelente	Muy baja	Azul
51-100	Buena	Baja	Verde
101-150	Mejorable	Elevada	Amarillo
>150	Deficiente	Muy elevada	Rojo

Tabla 2.2: Tramos cualitativos para los índices de calidad del aire y sus colores asociados.

2.10 Modelos de dispersión atmosférica

El objetivo de los modelos de simulación de la contaminación atmosférica es relacionar las emisiones de contaminantes con las concentraciones que alcanzan cuando llegan a los receptores, mediante algoritmos matemáticos que reproduzcan los efectos que la atmósfera induce sobre ellos. Se pueden clasificar en modelos estadísticos y modelos deterministas [62].

Los modelos estadísticos se basan en relaciones empíricas entre valores observados y registrados en el pasado, a lo largo de un periodo temporal extenso y son capaces de proporcionar una predicción, para las siguientes horas o días, de la concentración de contaminantes en una zona concreta, mediante una función estadística de valores medidos en ese momento y la correlación entre los valores de las medidas realizadas en el pasado y las tendencias de la concentración.

Los modelos deterministas se basan en descripciones matemáticas de los procesos atmosféricos para determinar el transporte, difusión, transformación y deposición superficial de los contaminantes, proporcionando una relación directa entre la fuente y el receptor. Estos modelos pueden clasificarse de acuerdo con el criterio de referencia (escala espacial, planteamiento de ecuaciones...) [8].

Por la forma en que se plantean las ecuaciones que describen el comportamiento de los contaminantes, se les puede clasificar en modelos eulerianos (usan un sistema de coordenadas fijo con respecto a la tierra) y modelos lagrangianos (usan un sistema de coordenadas que sigue el movimiento de la atmósfera).

Los modelos lagrangianos, a su vez, se pueden clasificar en modelos de trayectoria y modelos gaussianos, de acuerdo con la geometría del sistema de modelación. Estos modelos simulan la dispersión de los contaminantes como una columna de aire que se desplaza bajo la influencia de los vientos (modelos de trayectoria) o como una pluma de emisión, continua o discreta con paquetes llamados "puffs" (modelos gaussianos), donde se asume que las emisiones presentan una distribución normal o gaussiana con una concentración máxima en el centro de la pluma.

2.10.1. Modelos de dispersión recomendados por la EPA

La Agencia de Medio Ambiente de Estados Unidos[3] (EPA – *Environmental Protection Agency*) establece una serie de modelos de dispersión recomendados para abordar distintos problemas de calidad del aire:

- AERMOD [30]: Basado en la estructura (edificios) y conceptos de la turbulencia en la capa límite planetaria.
- BLP [78]: Modela emisiones industriales, especialmente de plantas de reducción de aluminio.
- CALINE3 [14]: Estima la contaminación atmosférica debida a carreteras, autopistas, etc.
- CAL3QHC/CAL3QHCR [33]: Basado en CALINE3. Calcula concentraciones en zonas de atascos. Requiere datos meteorológicos locales.
- CALPUFF [81]: Simula el efecto de las condiciones meteorológicas (con CALMET [80]) variando tiempo y espacio.
- CTDMPLUS [72]: Dispersión en terreno complejo para focos puntuales.
- OCD [41]: Determina el impacto de emisiones costeras y mar adentro. Necesita datos meteorológicos.

[3] EPA: https://www.epa.gov

CAPÍTULO 3
Minería de datos

En este capítulo se describe el proceso de minería de datos, así como las técnicas de regresión que serán utilizadas en este trabajo para predecir los niveles de contaminación.

3.1 Introducción

Hoy en día se generan datos constantemente: teléfonos móviles, redes sociales, información multimedia... Sensores y dispositivos liberan automáticamente información que necesita ser almacenada y procesada en tiempo real. Esta enorme cantidad de datos requiere nuevas herramientas y tecnologías para crear, manipular y gestionar los conjuntos de datos [82].

La minería de datos (*Data Mining*) es el proceso llevado a cabo para detectar información implícita en grandes conjuntos de datos, desconocida o previamente ignorada, utilizando análisis matemático (algoritmos) para deducir patrones y tendencias para crear, finalmente, un modelo de minería de datos [2][34]. Para que este proceso sea efectivo debe ser automático o semi-automático y el resultado debería ayudar a la toma de decisiones más seguras que reporten algún tipo de beneficio [42]. Se puede considerar, por tanto, como una colección de técnicas para inducir el conocimiento de una manera estructurada desde un gran conjunto de datos (descubrimiento de conocimiento en bases de datos, KDD) [5].

3.2 El proceso de minería de datos

Existen diferentes modelos o estándares que describen el proceso de la minería de datos, detallando la metodología que se debe seguir para realizar proyectos de este tipo y extraer conocimiento útil de los datos. La principal razón para establecer y usar un modelo es para organizar los proyectos de descubrimiento de conocimiento y minería de datos dentro de un marco común, ayudando a entender el proceso y proporcionando un camino a seguir [57].

En nuestro caso nos centraremos en la metodología descrita por el estándar CRISP-DM [27], puesto que es el más utilizado en la realización de proyectos de este tipo [3].

Según la metodología CRISP-DM, el proceso de minería de datos se puede definir en seis pasos:

- Definición del problema: Definir el ámbito y objetivos del problema y considerar formas de usar los datos para resolver el problema.

- Comprensión de los datos: Recolectar y conocer los datos, descubriendo información oculta para generar hipótesis.
- Preparación de los datos: Limpiar los datos, buscar correlaciones, determinar atributos...
- Generación de los modelos: Usar los conocimientos adquiridos en el paso anterior para definir y crear los modelos. El procesamiento de este modelo se denomina a menudo entrenamiento (aplicando un algoritmo concreto a los datos para extraer patrones).
- Validación de los modelos: Antes de implementar el modelo final, hay que comprobar que funciona correctamente.
- Implementación de los modelos: Producir los modelos finales.

3.3 Modelos de minería de datos

El conocimiento final que se ha extraído de los datos, constituye el modelo de los datos analizados. Los principales objetivos de estos modelos son predecir y describir. Los modelos predictivos estiman valores futuros o desconocidos. Los modelos descriptivos identifican patrones que explican o resumen los datos [34]. Dentro de los modelos predictivos podemos encontrar tareas de clasificación y regresión, mientras que dentro de los modelos descriptivos tenemos tareas de agrupamiento o *clustering*, reglas de asociación y correlaciones [42].

En concreto, en este trabajo se usarán dos de estas tareas, una descriptiva (correlación) y una predictiva (regresión). A continuación se verán estas dos tareas con más detalle.

3.3.1. Modelos descriptivos: Correlaciones

Las correlaciones son tareas descriptivas, usadas para determinar la similitud entre los valores de dos variables numéricas [42].

Para medir la correlación existe una fórmula estándar: el coeficiente de correlación r (coeficiente de Pearson [70]). r es un valor real comprendido en el intervalo $[-1, 1]$, de modo que:

- Si $r = 1$, existe correlación positiva perfecta, es decir cuando una aumenta, la otra también lo hace. Lo mismo sucede cuando disminuyen.
- Si $0 < r < 1$, existe correlación positiva.
- Si $r = 0$, no existe relación lineal entre las variables.
- Si $-1 < r < 0$, existe correlación negativa.
- Si $r = -1$, existe correlación negativa perfecta, lo que quiere decir que cuando una variable aumenta, la otra disminuye y viceversa en el caso contrario.

El análisis de estas correlaciones resulta muy útil para establecer relaciones y reglas entre las variables.

En el capítulo 5 veremos la utilización de correlaciones aplicadas a los datos de contaminación, meteorológicos y de tráfico. De igual manera, veremos el uso del coeficiente de correlación a la hora de evaluar el modelo de predicción, en el capítulo 6.

3.3.2. Modelos predictivos: Regresión

El análisis predictivo es un área de la minería de datos que agrupa una serie de técnicas estadísticas de modelización y aprendizaje automático[1] para analizar datos actuales e históricos con el objetivo de realizar predicciones, identificando relaciones entre las variables [40].

La regresión es una tarea del análisis predictivo, donde el objetivo es explicar el comportamiento de una variable (variable respuesta) a partir del conocimiento de otras (variables explicativas), aprendiendo para ello una función real y siendo el valor a predecir numérico. Si solo se dispone de una variable explicativa, se llama regresión simple. Si se dispone de varias, entonces hablamos de regresión múltiple. El objetivo en este tipo de técnicas es minimizar el error cometido entre el valor predicho y el valor real [42].

Aprendizaje supervisado y semi-supervisado

Las técnicas de regresión se engloban dentro del aprendizaje automático en la categoría de aprendizaje supervisado o semi-supervisado, es decir aquel en el hay datos etiquetados.

En este tipo de aprendizaje, los datos de entrada son llamados *training data* (datos de entrenamiento) y el modelo se prepara a través de un entrenamiento donde debe aprender la estructura de los datos y organizarlos para realizar predicciones sobre un conjunto de test (prueba) [20].

Técnicas de regresión

En el capítulo 6, realizaremos experimentos utilizando diferentes técnicas de regresión con el fin de identificar aquella que es capaz de realizar la mejor predicción de los niveles de contaminación. En concreto se probarán las siguientes técnicas:

- ***Linear Regression*:** Permite determinar el grado de dependencia de las series de valores X e Y, prediciendo el valor y estimado que se obtendría para un valor x que no esté en la distribución [5].
- ***Quantile regression*:** *Quantile regression* extiende el modelo de regresión a cuantiles condicionales de la variable respuesta, transforma una función de distribución condicional en una función cuantil condicional, separándola en segmentos. Estos segmentos describen la distribución acumulativa de una variable dependiente condicional, dada una variable explicativa x_i con el uso de cuantiles [28].

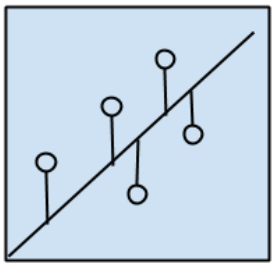

Figura 3.1: Esquema gráfico de las técnica de regresión. Fuente: [20].

[1] El aprendizaje automático (*Machine Learning*) investiga como los computadores pueden aprender y mejorar en base a los datos, reconociendo complejos patrones y tomando decisiones basándose en ellos [40].

- ***K nearest neighbours* (IBK):** En el algoritmo kNN la ponderación permite a los vecinos más cercanos al punto tener más influencia en el valor a predecir. IBK selecciona el valor más adecuado basándose en validación cruzada , realizando ponderación a la distancia usando una medida simple de distancia para encontrar la instancia más cercana a la instancia de prueba dada. Si más de una instancia tienen la misma distancia a la instancia de prueba, la primera encontrada es la usada por el algoritmo [89].

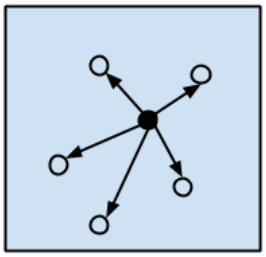

Figura 3.2: Esquema gráfico de las técnica de kNN. Fuente: [20].

- **Arboles de decisión (M5P):** Un árbol de decisión es similar a un diagrama de flujo en el que cada nodo interno representa una prueba o acción en un atributo, cada rama representa el resultado de la prueba y cada hoja representa el resultado final. Los caminos de la raíz a la hoja representan las reglas o decisiones. M5P combina un árbol de decisión convencional con la posibilidad de incluir funciones de regresión lineal en los nodos [19].

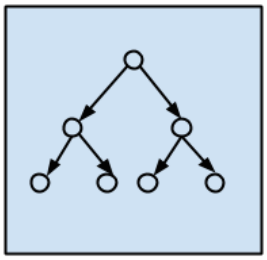

Figura 3.3: Esquema gráfico de las técnica de árboles de decisión. Fuente: [20].

- ***Random Forest*:** *Random Forest* es la aplicación de *bagging* (combinación) de multitud de árboles de decisión, en el que se selecciona para cada árbol un vector de atributos aleatorio [59].

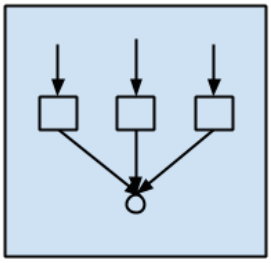

Figura 3.4: Esquema gráfico de las técnica de *Random Forest*. Fuente: [20].

En el capítulo 8, aplicaremos los modelos entrenados con la técnica *Random Forest* para predecir los valores de los contaminantes y poder mostrarlos posteriormente en la aplicación web.

3.4 Tecnologías utilizadas

Todos los experimentos relacionados con técnicas de regresión realizados en este trabajo se han llevado a cabo utilizando el lenguaje estadístico R [74] en su versión software para Windows, RStudio[2]. En concreto se han usado las siguientes librerías:

- RWeka: Interfaz R para Weka[3] [4].
- caret: Funciones para entrenar y visualizar modelos de regresión y clasificación[5].
- quantreg: Librería para la utilización de modelos de *Quantile Regression*[6].
- randomForest: Clasificación y regresión basado en arboles de decisión usando variables de entrada aleatorias[7].

3.5 Trabajos relacionados

Las técnicas de minería de datos han sido anteriormente utilizadas para predecir los niveles de contaminación.

Un trabajo fundamental en este área con Redes Neuronales es [94]. Las Redes Neuronales se utilizan con bastante frecuencia en este campo, un resumen de estas aproximaciones se puede ver en [55].

Un trabajo más relacionado es [53]. Aquí, los autores proponen un sistema para predecir los volúmenes de tráfico, las emisiones de los vehículos y la dispersión atmosférica en un área urbana. El artículo compara las predicciones en las concentraciones de NO y NO_2 con los resultados de una red de monitores de calidad del aire. Los resultados para las predicciones son mejores para dos estaciones suburbanas comparadas con dos estaciones urbanas.

Nuestra comparación de modelos de regresión obtiene resultados similares a las conclusiones presentadas en el trabajo de [85]. En este estudio, se usa el Análisis de Componentes Principales (PCA) para identificar las fuentes de contaminación. De los componentes extraídos, se generan tres modelos para predecir la calidad de aire urbano en Lucknow (India), junto con datos meteorológicos y de contaminación de un periodo de cinco años.

La dirección del viento ha sido raramente incorporada en modelos espaciales (*Land-Use Regression Models*, LUR). [45] identifica 25 estudios de LUR y solo dos incorporan la dirección del viento en sus modelos predictivos. [12] estudia el uso de flujos de viento para mejorar la predicción de contaminación del aire en Toronto. Los flujos de viento se extraen desde 38 estaciones y son especialmente eficientes para mejorar la predicción del NO_2. Otra aproximación es [87]. Aquí, los autores aplican LUR integrando la velocidad del viento, la dirección del viento y la nubosidad para estimar concentraciones horarias de NO y NO_2.

[2] RStudio: https://www.rstudio.com/.

[3] Weka es una herramienta para minería de datos escrita en java que dispone de una colección de librerías de aprendizaje automático para clasificación, regresión, agrupamiento, etc.

[4] RWeka: https://cran.r-project.org/web/packages/RWeka/index.html

[5] caret: https://cran.r-project.org/web/packages/caret/index.html

[6] quantreg: https://cran.r-project.org/web/packages/quantreg/index.html

[7] randomForest: https://cran.r-project.org/web/packages/randomForest/index.html

CAPÍTULO 4
Interpolación espacial

En este capítulo se describe la interpolación espacial, usada en este trabajo para extrapolar los datos de los contaminantes a toda la ciudad.

4.1 Introducción

Debido al alto coste de los sensores para la recolección de ciertos tipos de datos, como los meteorológicos o los de contaminación, la red de estos sensores suele ser pequeña y no estar homogéneamente repartida, dejando muchos puntos sin la posibilidad de disponer de datos reales. La interpolación espacial es el proceso por el cual se estiman los valores desconocidos de cualquier punto geográfico, utilizando puntos con valores conocidos [67]. En la figura 4.1 podemos ver un esquema de este tipo de proceso.

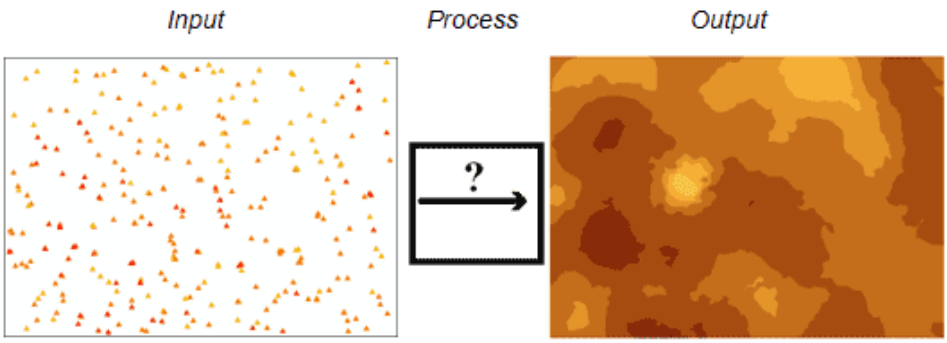

Figura 4.1: Proceso de interpolación. Estimación de los valores en puntos donde no hay valores a partir de valores conocidos en otros puntos. Fuente: [21].

4.2 Tipos de interpolación

Las técnicas de interpolación se pueden agrupar en dos categorías principales: deterministas y geoestadísticos.

Las ténicas de interpolación deterministas crean superficies a partir de puntos medidos, ya sea basado en el grado de similitud (por ejemplo, IDW) o en el grado de suavizado (por ejemplo, CRS) [47].

Las técnicas geoestadísticas (como Kriging), utilizan propiedades estadísticas de los puntos con valores, cuantificando la correlación entre todos los puntos de alrededor de la ubicación a predecir [17].

A su vez, los métodos pueden dividirse en métodos globales, si utilizan toda la muestra de datos para estimar el valor en cada nuevo punto, y métodos locales, si utilizan solo los puntos de muestreo más cercanos al punto a predecir [4].

4.2.1. Métodos globales

Los métodos globales asumen que la variable a predecir depende de otras variables de apoyo. Dependiendo del tipo de variable de apoyo se pueden dar dos casos, los métodos de clasificación, si la variable de apoyo es cualitativa o los de regresión, si es cuantitativa.

4.2.2. Métodos locales

Los métodos locales utilizan los puntos más cercanos (conjunto de interpolación) al punto de interpolación, asumiendo autocorrelación espacial.

Métodos locales basados en medias ponderadas

En este tipo de métodos, los valores se calculan como una media ponderadas de los valores del conjunto cercano. Para esto se toman dos decisiones:

1. Qué puntos van a formar parte del conjunto de interpolación.
   - Distancia menor que un umbral r.
   - Los n puntos más cercanos.
2. Cuál será el método de interpolación.
   - Asignar el valor del vecino más próximo.
   - Asignar la media de los puntos del conjunto de interpolación.
   - Asignar media ponderada por la distancia (IDW) de los puntos del conjunto de interpolación.
   - Utilizar *kriggeado*. Este método se vale de la información del semivariograma para obtener factores de ponderación optimizados.

Interpolación local por *splines*

El método de *splines* es similar a una interpolación global por regresión, pero a nivel local, ajustando ecuaciones polinómicas siendo las variables independientes X e Y (longitud y latitud). La ventaja de este método es el de poder generar valores máximos y mínimos, imposibles de obtener mediante medias ponderadas.

Interpolación local mediante TIN

La interpolación mediante TIN (Redes Triangulares de Triángulos) se generan tratando de conseguir triángulos que maximicen la relación en el área. El conjunto de todos los triángulos se denomina conjunto convexo.

4.3 Visualización de resultados

Los resultados de estas técnicas normalmente se suelen representar en una imagen o mapa, con zonas de diferentes colores o tonalidades según el valor o la tendencia de los valores. En la figura 4.2 se pueden ver los diferentes resultados que pueden producir algunos de estos métodos aplicados a datos de temperatura, obtenidos en los experimentos de [11].

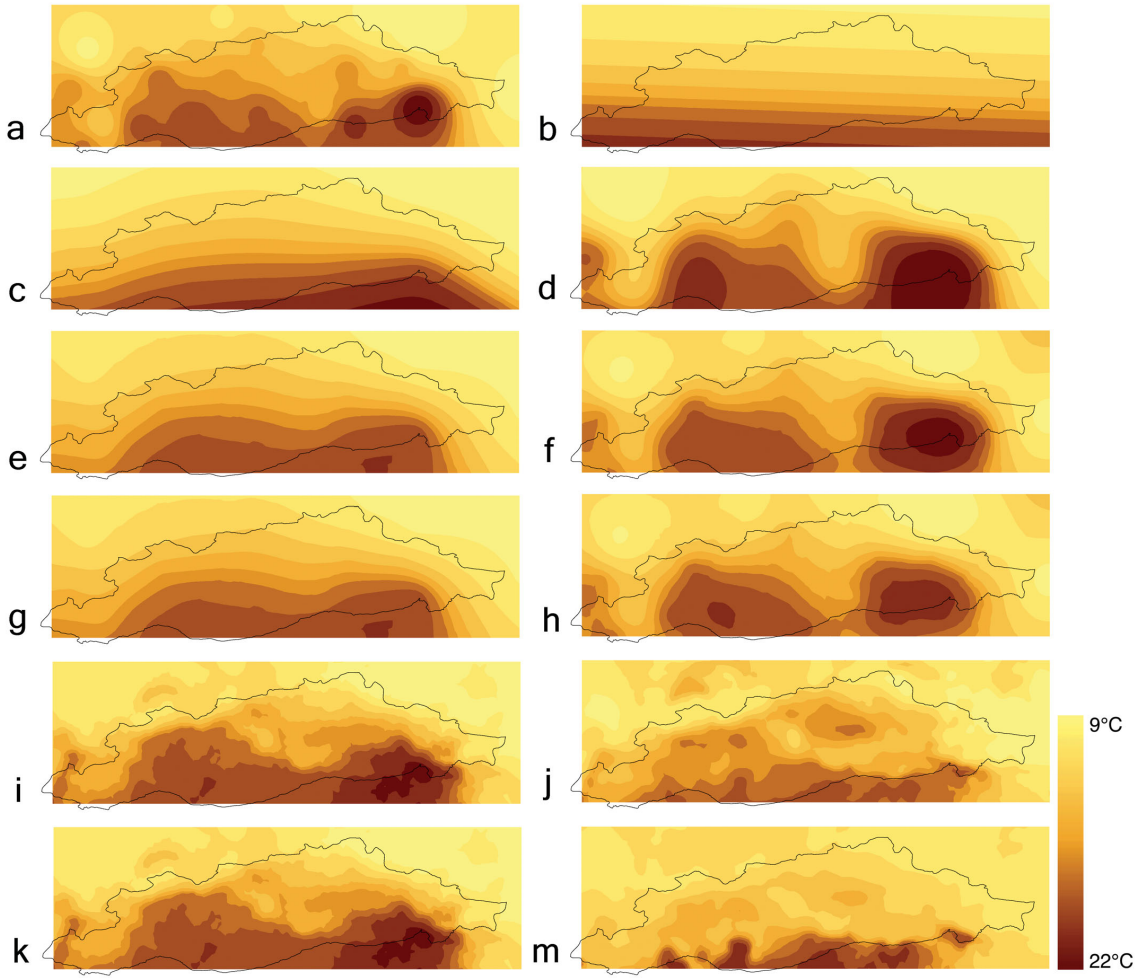

Figura 4.2: Interpolación de datos de temperatura en una región de Turquía. *(a) Inverse Distance Weighted, (b) Global Polynomial Interpolation, (c) Local Polynomial Interpolation, (d) Completely Regularized Spline, (e) Ordinary Kriging, (f) Simple Kriging, (g) Universal Kriging, (h) Disjunctive Kriging, (i) Ordinary CoKriging, (j) Simple CoKriging, (k) Universal CoKriging, (m) Disjunctive CoKriging*. Fuente: [11].

4.4 Tecnologías utilizadas

Los experimentos con interpolación, al igual que los de predicción, se han realizado con el lenguaje R. En este caso las librerías utilizadas son las siguientes:

- geosphere: Funciones de trigonometría esférica para aplicaciones geográficas[1].
- geoR: Ánalisis geoestadístico (incluye Kriging)[2].

[1] geosphere: https://cran.r-project.org/web/packages/geosphere/index.html
[2] geoR: https://cran.r-project.org/web/packages/geoR/index.html

- ggmap: Colección de funciones para visualizar datos espaciales y modelos en mapas estáticos desde diferentes fuentes[3].

4.5 Trabajos relacionados

En el capítulo 7 solo veremos dos de estas técnicas conocidas (IDW y Kriging), sin embargo, algunos trabajos ya han empleado la interpolación anteriormente para estimar el nivel de los contaminantes en zonas sin datos.

En [22] los autores comparan las técnicas IDW y Kriging con datos de particulas en suspensión (PM_{10}) en Madrid, realizando un minucioso estudio entre las diferencias aportadas por ambos métodos. Sus resultados se asemejan a los que veremos en este trabajo, puesto que Kriging les proporciona mejores resultados que IDW.

El método RIO se presenta en [46] como un modelo de interpolación para la polución del aire. El método usa un parámetro β que ofrece flexibilidad a la hora de ponderar entre LUR [45] y los niveles de contaminación. Los experimentos con O_3, NO_2 y PM_{10} con un procedimiento de validación cruzada, muestran que RIO produce mejores resultados comparado con IDW y Ordinary Kriging.

En [63] los autores comparan LUR y Universal Kriging (UK). En sus experimentos con modelos predictivos para NO_x en Los Ángeles (EEUU), la interpolación con UK supera consistentemente a LUR. La idea básica que sigue LUR es calcular una regresión lineal múltiple en capas de mapas de bits o vectoriales que contienen información del uso de la tierra, tales como distancias de las carreteras, aeropuertos, redes hidrográficas, población, tráfico, etc. En [26] se generan dos modelos usando LUR y técnicas de aprendizaje automático (*Multi Layer Perceptron* y *Random Forest*), comparándolos con herramientas lineales. Sus experimentos demuestran que las técnicas de aprendizaje automático pueden mejorar los resultados obtenidos por los modelos de interpolación espacial.

De hecho, el uso de modelos de aprendizaje automático y ténicas de interpolación espacial conjuntamente han producido buenos resultados en muchos problemas reales [52], superando las dificultades que por separado pueden tener ambos enfoques.

[3] ggmap: https://cran.r-project.org/web/packages/ggmap/

CAPÍTULO 5

Análisis y preparación de los datos

Seleccionar de forma correcta las fuentes de datos, estudiar en profundidad estos datos con el fin de extraer nuevo conocimiento o identificar y eliminar inconsistencias, discrepancias y errores en los mismos, para mejorar la calidad, son etapas de suma importancia en un proyecto de *data mining* como el que aquí se describe.

En esta sección se describirá el proceso seguido para seleccionar, analizar y preparar los datos y parámetros para su posterior utilización en los experimentos de este trabajo.

5.1 Introducción

Valencia dispone de seis estaciones de medición de los niveles de contaminación distribuidas a lo largo de toda la ciudad. Sus ubicaciones exactas pueden verse en la Figura 5.1.

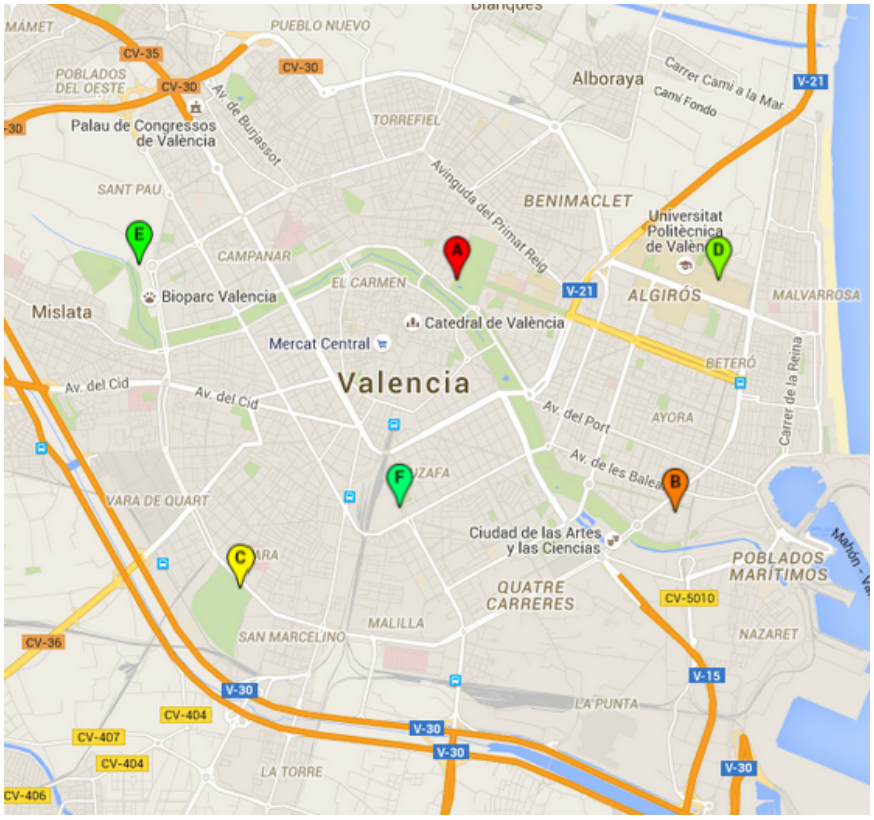

Figura 5.1: Ubicación de las estaciones de medición de contaminación en la ciudad de Valencia.

En la tabla 5.1 podemos ver las características de cada una de estas estaciones disponibles.

Estación	*Zona*	*Emisión*	*Situación*
Bulevar	Urbana	Tráfico	Situada en el parking del Cementerio General, en el Bulevar Sur. Sin calles o edificios cercanos. Pocos puntos de tráfico cercanos. Tiene al suroeste la autovía V-30.
Francia	Urbana	Tráfico	Situada en la Avenida de Francia, rodeada de calles y edificios, con el cauce del rio al sur y el puerto de Valencia a 1km.
Molí	Suburbana	Tráfico	Situada entre la Avenida Pío Baroja y la Avenida General Avilés. Sus únicos puntos de tráfico cercanos se sitúan al este. Tiene al sur el Bioparc y al oeste el parque de la canaleta.
Pista	Urbana	Tráfico	Situada en la calle Filipinas y rodeada de calles y edificios. El centro de Valencia está a menos de 1km de la estación. Tiene al oeste las vías del tren de la Estación del Norte y los barrios de Ruzafa y Malilla al este y sur, respectivamente.
UPV	Suburbana	Fondo	Situada dentro de la Universidad Politécnica de Valencia, sin apenas puntos de tráfico cercanos.
Viveros	Urbana	Fondo	Situada dentro de los jardines de Viveros. Sus únicos puntos de tráfico cercanos provienen de la Calle Alboraya al norte y este. El cauce del rio se ubica a 200 metros al sur.

Tabla 5.1: Características de las estaciones de medición de contaminación

El portal de datos abiertos de la Generalitat Valenciana[1] proporciona los datos de los niveles de contaminación medidos por cada una de las estaciones, con un retraso de tres horas. Esta demora en el servicio es debido al tiempo que requiere la necesaria validación de los contaminantes de menor dimensión, como las partículas menores de 2,5μm ($PM_{2,5}$). Por este motivo, los datos de contaminación nunca son publicados en tiempo real y por tanto, una alerta de un alto nivel en los contaminantes llegaría también con retraso.

Dado que los datos de contaminación nunca estarán disponibles en tiempo real, se requiere la utilización de otros parámetros que puedan estar relacionados con los contaminantes, para poder estimar los valores de contaminación con antelación a su publicación, a partir de ellos.

5.2 Selección de contaminantes

Las seis estaciones de medición de contaminación de Valencia evalúan diferentes contaminantes pero los datos muchas veces no llegan a ser validados, por tanto, hay grandes espacios de tiempo carentes de información, lo que dificulta su estudio y utilización. En nuestro caso, se ha decidido estudiar un periodo de tiempo de tres años (comprendido entre enero de 2013 y diciembre de 2015), seleccionando aquellos contaminantes que provienen total o parcialmente de emisiones antropogénicas, como la combustión de los motores de los vehículos (por ser la fuente de emisión más perjudicial en las ciudades),

[1]Portal de datos abiertos de la GVA, Calidad ambiental: http://www.habitatge.gva.es/web/calidad-ambiental/datos-on-line

presentes en las mediciones de las seis estaciones y con un porcentaje de datos válidos superior al 60 %.

	Año	*CO*	*NO*	NO_2	$PM_{2,5}$	PM_{10}	SO_2	O_3
Bulevar	2013	25.38	72.37	72.37			82.25	97.26
	2014	89.21	89.21	89.21			89.49	98.69
	2015	8.46	86.46	86.46			90.23	97.66
	TOTAL	**41.02**	**82.68**	**82.68**	**0**	**0**	**87.32**	**97.87**
Francia	2013	90.7	79.95	79.95	93.43	93.37	84.89	97.43
	2014	81.95	81.11	81.11	92.6	92.6	76.63	92.6
	2015	90.13	84.19	84.19	77.64	77.64	85.11	97.04
	TOTAL	**87.59**	**81.75**	**81.75**	**87.89**	**87.87**	**82.21**	**95.69**
Molí	2013	48.13	79.59	79.59	72.08	72.08	94.16	96.61
	2014	88.25	89.18	88.84	96.63	96.63	65.66	94.17
	2015	74.09	86.39	86.27	84.91	84.91	96.1	95.76
	TOTAL	**70.16**	**85.05**	**84.90**	**84.54**	**84.54**	**85.31**	**95.51**
Pista	2013	73.56	69.43	69.43		99.79	71.51	99.27
	2014	96.28	99.69	99.69	99.34	99.34	89.83	97.23
	2015	86.73	88.97	88.97	64.23	97.49	86.28	89.68
	TOTAL	**85.52**	**86.03**	**86.03**	**54.52**	**98.87**	**82.54**	**95.39**
UPV	2013		84.26	84.26	97.31	97.31	68.2	82.93
	2014		96.68	96.68	97.93	97.93	88.65	93.42
	2015		92.82	92.82	97.9	97.9	83.36	91.07
	TOTAL	**0**	**91.25**	**91.25**	**97.71**	**97.71**	**80.07**	**89.14**
Viveros	2013	88.39	63.25	63.25			74.24	85.53
	2014		92.12	92.12			91.02	97.19
	2015	29.46	81.9	81.9			83.74	90.41
	TOTAL	**39.28**	**79.09**	**79.09**	**0**	**0**	**83.00**	**91.04**

Tabla 5.2: Porcentaje de datos completos por contaminante, estación y año para las seis estaciones de contaminación de Valencia. En rojo aquellos contaminantes que no llegan al porcentaje mínimo.

En la Tabla 5.2 podemos ver el porcentaje de datos completos disponibles para los años 2013, 2014 y 2015, para siete contaminantes relacionados de alguna manera con las emisiones antropogénicas, en las seis estaciones disponibles. Como se puede observar, las partículas en suspensión (tanto PM_{10} como $PM_{2,5}$) no están disponibles en todas las estaciones (las estaciones de Bulevar y Viveros no disponen de mediciones). De la misma manera, no existen datos de monóxido de carbono (CO) en la estación de la UPV y no hay suficientes datos en la estación de Viveros a partir de 2014.

Por tanto, en este trabajo nos hemos centrado en los siguientes cuatro contaminantes [60][48][7], cuya cantidad de datos es suficiente para realizar el estudio:

- **NO_x (Óxidos de nitrógeno):** El óxido nítrico (NO) y el dióxido de nitrógeno (NO_2) se suelen considerar en conjunto con la denominación de NO_x. Son contaminantes primarios de mucha trascendencia en los problemas de contaminación. NO_x tiene una vida corta y se oxida rápidamente a NO_3. Tiene una gran relevancia en la formación del smog fotoquímico, del nitrato de peroxiacetilo (PAN) e influye en las reacciones de formación y destrucción del ozono, tanto troposférico como estratosférico, así como en el fenómeno de la lluvia ácida. En concentraciones altas produce daños a la salud y a las plantas y corroe tejidos y materiales diversos. Las actividades humanas que los producen son, principalmente, las combustiones realizadas a altas temperaturas. Más de la mitad de los gases de este grupo emitidos en España proceden del transporte.

  - **NO (Monóxido de nitrógeno):** Gas incoloro, inodoro, no inflamable y tóxico por su capacidad para oxidarse a NO_2. Se ha escogido para este estudio el monóxido de nitrógeno por varios motivos, pese a su escasa duración en

la atmósfera en presencia de luz solar. Una de estas razones es que su origen proviene en gran medida de los motores de los vehículos, por lo que es un parámetro adecuado para relacionarlo directamente con el tráfico. Por otra parte, el monóxido de nitrógeno es un compuesto altamente inestable, que en la atmósfera reacciona rápidamente originando dióxido de nitrógeno. Esta inestabilidad convierte al monóxido de nitrógeno en un radical, es decir, una molécula con un alto poder reactivo, cuyos efectos en el organismo son la alteración del ADN, los lípidos y las proteínas. Este tipo de cambios derivan a medio y largo plazo en una mayor probabilidad de sufrir cáncer.

- **NO_2 (Dióxido de nitrógeno):** Gas de tonalidad rojiza, de fuerte olor, no inflamable, muy corrosivo y tóxico (cuatro veces más que el NO). Se ha seleccionado para su estudio el dióxido de nitrógeno por estar íntimamente relacionado con el monóxido de nitrógeno, y por sus efectos en el organismo. El dióxido de nitrógeno es un contaminante que no se genera de forma directa, puesto que su presencia en la atmósfera se debe a la oxidación del monóxido de nitrógeno. Se trata de un compuesto que en presencia de humedad deriva en ácido nítrico, y está demostrado que su inhalación, aún en bajas concentraciones, acaba originando degradación del tejido pulmonar, así como en una disminución de la eficacia del sistema inmunitario, especialmente en niños.

- **SO_2 (Dióxido de azufre):** Gas bastante estable, incoloro, no inflamable y muy soluble en agua, con un olor fuerte e irritante. El dióxido de azufre es un gas tóxico producido principalmente por la actividad volcánica y cuya fuente de emisión principal, debida a la actividad por parte del hombre, son la industria y la combustión de carburantes. Las emisiones de SO_2 están directamente relacionadas con los procesos de acidificación, ya que al oxidarse produce ácido sulfúrico (H_2SO_4), uno de los causantes de la lluvia acida. Inhalar este contaminante produce irritación de ojos, mucosas y piel, enfermedades respiratorias y muerte prematura. También produce efectos sobre la vegetación, favoreciendo la necrosis.

- **O_3 (Ozono troposférico):** Gas incoloro con gran poder oxidante. Este ozono es el existente en la troposfera, la región inferior de la atmósfera terrestre. A este nivel, la presencia de ozono a elevadas concentraciones tiene efectos perjudiciales sobre la salud y el medio ambiente en general, reduce significativamente la función pulmonar e induce a inflamaciones respiratorias. Absorbe radiación infrarroja potenciando el efecto invernadero y a concentraciones elevadas es el componente más dañino del smog fotoquímico. El Ozono no es liberado directamente de una fuente específica, sino que es generado por la acción de la luz solar sobre los NO_x y los compuestos orgánicos volátiles (VOC) del aire, además de generarse a partir de los procesos de combustión (transporte y generación de energía eléctrica). Los niveles de ozono varían a lo largo del día, dependiendo de la intensidad del tráfico, de la actividad industrial y de la intensidad de la luz solar. En el Mediterráneo es en verano cuando se dan condiciones meteorológicas favorables para su formación, puesto que su origen se ve favorecido en situaciones estacionarias de altas presiones asociadas a una fuerte insolación y vientos débiles.

5.3 Fuentes de datos

5.3.1. Datos de contaminación

Para este trabajo se han utilizado datos históricos horarios de contaminación, obtenidos del portal de datos abiertos de la Generalitat Valenciana.

Los datos históricos de contaminación disponibles en este portal se ofrecen en dos formatos, como medias horarias y como medias diarias. En este caso, se han utilizado las medias horarias. Los datos están clasificados por año y estación y descargables en archivos de texto (.txt). Cada uno de estos archivos dispone de los siguientes datos:

- Fecha, en formato DD/MM/AAAA.
- Hora, en formato HH.
- Contaminantes ($PM_{2,5}$, PM10, SO_2, CO, etc.), dependiendo de la estación, en $\mu g/m^3$ o mg/m^3 dependiendo del contaminante.
- Velocidad del viento (únicamente en las estaciones de Avenida de Francia y Pista de Silla), en m/s.
- Dirección del viento (únicamente en las estaciones de Avenida de Francia y Pista de Silla), en grados.
- Ruido (únicamente en las estaciones de la Pista de Silla y Viveros), en dBA.
- Humedad relativa (únicamente en la estación de la Pista de Silla), en porcentaje.
- Presión (únicamente en la estación de la Pista de Silla), en mb.
- Radiación Solar (únicamente en la estación de la Pista de Silla) en W/m^2.

Para este trabajo se han utilizado los datos de fecha, hora y los cuatro contaminantes seleccionados anteriormente (NO, NO_2, O_3, SO_2) en cada estación.

5.3.2. Datos de la intensidad del tráfico

En los países desarrollados, la fuente principal de contaminación son los vehículos a motor y la industria. Los vehículos liberan grandes cantidades de óxidos de nitrógeno, óxidos de carbono, hidrocarburos y partículas al quemar la gasolina y el gasóleo. Por esta razón, es útil medir el nivel de tráfico en una ciudad, para poder estudiar la relación existente entre la cantidad de vehículos circulando por una zona y la contaminación total medida.

La ciudad de Valencia dispone de una red de 1231 sensores (espiras electromagnéticas) que proporcionan datos de la intensidad de tráfico (Vehículos/hora) en tiempo real. En la figura 5.2 podemos ver la distribución de estas espiras en la ciudad.

Se han utilizado datos históricos horarios de la intensidad de tráfico medido por las espiras electromagnéticas. Esta información se puede encontrar en el portal de datos abiertos del Ayuntamiento de Valencia[2] en tiempo real.

En nuestro caso, los históricos sobre datos de tráfico fueron proporcionados por la jefatura de tráfico de la ciudad de Valencia. Estos datos se encuentran en un solo archivo en formato de bases de datos de Microsoft Access (.mdb), donde cada tabla contiene los datos de un mes para cada uno de los años estudiados. En concreto, cada tabla se compone de los siguientes datos:

- Punto de medida (identificador de la espira), en formato numérico (entero).

[2]Portal de datos abiertos del Ayuntamiento de Valencia, Intensidad de tráfico: http://gobiernoabierto.valencia.es/es/dataset/?id=puntos-medida-trafico-espiras-electromagneticas

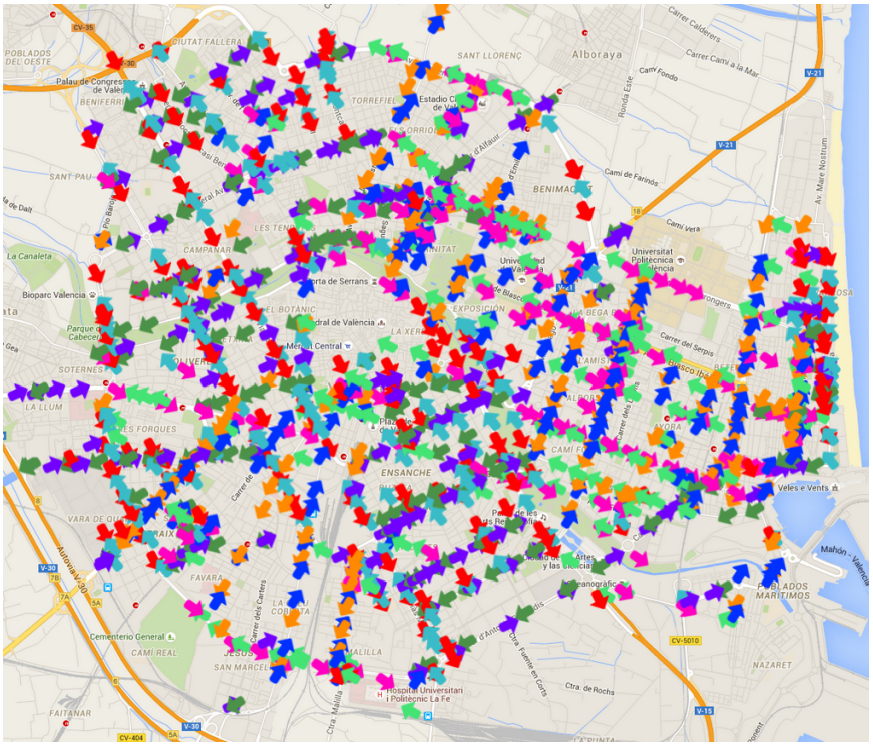

Figura 5.2: Ubicación de las espiras electromagnéticas de medición de la intensidad de tráfico en la ciudad de Valencia (las flechas indican el sentido de la calle, dirección en la que se mueve el tráfico).

- Fecha y hora, en formato D/M/AAAA H:MM:SS.
- Intensidad horaria, en formato numérico (entero).
- Fiabilidad del dato, en porcentaje.
- Ocupación de la vía, en formato numérico (entero).
- Fiabilidad del dato de ocupación de la vía, en porcentaje.
- Velocidad de los vehículos en la vía, en formato numérico (entero).
- Fiabilidad del dato de velocidad de los vehículos en la vía, en porcentaje.
- Cola, dato binario {0,1}.

De cada una de estas tablas se han utilizado los datos del identificador del punto de medida, la fecha y hora y la intensidad horaria.

Además de estos datos, se necesitaba la posición real de las espiras (ubicación), tanto para posicionarlas y visualizarlas en el mapa, como para realizar algunos cálculos necesarios en el trabajo. Por esta razón, se descargó un archivo de datos de la intensidad de tráfico en los puntos de medida en formato csv, disponible en tiempo real en la web del Ayuntamiento. Este archivo contiene los siguientes datos:

- Coordenadas de las espiras, en formato EPSG:25830 (ETRS89 / UTM 30N)[3].
- Ángulo de la espira (sentido de la calle), en grados.

[3]EPSG:25830 http://spatialreference.org/ref/epsg/etrs89-utm-zone-30n/

- Fecha de actualización de los datos, en formato AAAA-MM-DD.
- Hora de actualización de los datos, en formato HH:MM:SS.
- Identificador del punto de medida, en formato numérico (entero).
- Intensidad horaria, en formato numérico (entero).

De estos datos utilizamos el identificador de las espiras y las coordenadas de las mismas.

5.3.3. Datos meteorológicos

Las circunstancias climatológicas influyen de modo determinante en la distribución de la contaminación atmosférica, influenciando la generación y dispersión de la contaminación. El conocimiento de todos estos factores a nivel micrometeorológico[4], es indispensable para el estudio de los niveles de contaminación de los núcleos urbanos e industriales. Los fuertes vientos, por ejemplo, pueden dispersar los contaminantes lejos de su punto de emisión.

Por este motivo, hemos recopilado la observación meteorológica horaria de la ciudad de Valencia del portal de datos abiertos de la Generalitat Valenciana para el periodo de tres años. Los datos disponibles en esta web, descargados en formato txt, recogen los datos horarios de la observación meteorológica. En concreto, están disponibles con los siguientes datos:

- Año, en formato AAAA (numérico, entero).
- Mes, en formato M (numérico, entero).
- Día, en formato D (numérico, entero).
- Hora, en formato H (numérico, entero).
- Día de la semana (Lunes, Martes...).
- Velocidad del viento máxima, en m/s.
- Velocidad del viento, en m/s.
- Dirección del viento, en grados.
- Temperatura, en grados Celsius.
- Humedad relativa, en porcentaje.
- Presión, en mb.
- Precipitación, en l/m^2.

En concreto, hemos utilizado los siguientes parámetros meteorológicos, relacionados con la contaminación:

[4] Fenómenos meteorológicos que ocurren en distancias del orden de los 10km

- Temperatura: En una situación habitual de la atmósfera, la temperatura desciende con la altitud lo que favorece que ascienda el aire más caliente (menos denso) y arrastre a los contaminantes hacia arriba. En una situación de inversión térmica una capa de aire más cálido se sitúa sobre el aire superficial más frío e impide la ascensión de este último (más denso), por lo que la contaminación queda encerrada y en aumento. Se ha escogido como parámetro la temperatura por su íntima relación con la concentración o dispersión de los parámetros contaminantes.
- Humedad relativa: La humedad es un factor climatológico a tener en cuenta, ya que en su presencia el dióxido de nitrógeno deriva en ácido nítrico, perjudicial para la salud humana.
- Presión: El viento se genera a causa de diferencias en la presión de la atmósfera (la presión es el peso de la atmósfera en un momento determinado).
- Velocidad del viento: El viento puede dispersar los agentes contaminantes y transportarlos lejos de su punto de emisión. Es por esto, uno de los factores climatológicos más importantes a tener en cuenta para el cálculo de la contaminación existente.
- Precipitaciones: Las precipitaciones arrastran los contaminantes y disuelven sustancias y gases, lo que hace que sea necesario tener en cuenta este parámetro para estudiar su relación con la contaminación.

5.4 Limpieza y preparación de los datos

Para cada una de las seis estaciones de contaminación de Valencia hemos creado un conjunto de datos horarios con el nivel de los contaminantes y los parámetros que pueden afectarles, con los datos del periodo de tres años mencionado anteriormente. En concreto, hemos extraído los siguientes datos para cada una de las estaciones:

- **Nivel de los contaminantes ($\mu g/m^3$):** NO, NO_2, O_3, SO_2
- **Condiciones meteorológicas:** Temperatura (Grados Celsius), humedad relativa (Porcentaje), presión (mb), velocidad del viento (m/s), precipitaciones (mm/h).
- **Características temporales:** Año, mes, día del mes, día de la semana, hora.
- **Intensidad del tráfico:** Nivel de tráfico alrededor de la estación (suma del tráfico total en un radio de 1km).

En cada uno de estos conjuntos de datos se han unido los datos procedentes de las diferentes fuentes de datos usadas. Puesto que se ha creado un conjunto de datos por estación, con los datos de los tres años, en cada archivo (csv) se han juntado los datos de los contaminantes de la estación correspondiente.

Respecto a los datos históricos de contaminación, cabe destacar la gran cantidad de datos faltantes (meses enteros en algunas estaciones). Igualmente, entre los datos de la Avenida de Francia y Molí del Sol de 2014, se podían encontrar comas en los decimales (siendo el punto el símbolo utilizado para el resto de los datos), lo que ha sido sinónimo de más de un error difícil de identificar en R.

Como se ha visto en las fuentes de datos, la fecha y la hora son proporcionadas en un formato diferente en cada fuente. Antes de poder unir los datos en un solo archivo por estación, ha sido necesario realizar conversiones en las fechas y horas, ya que la intención

al juntar los datos horarios es unirlos haciéndolos coincidir por las fechas. Esto se ha conseguido mediante las funciones format[5] y as.Date[6] de R.

Para calcular la suma del tráfico alrededor de cada estación, se ha creado una matriz de distancias entre las estaciones de medición de contaminación y los sensores de tráfico. Para este propósito, se ha hecho uso de la fórmula de Haversine [84], que calcula la distancia de círculo máximo entre dos puntos de una esfera sabiendo su longitud y latitud. En concreto, para cualquier par de puntos sobre una esfera:

$$hav(\frac{d}{r}) = hav(\phi_2 - \phi_1) + \cos(\phi_1)\cos(\phi_2)hav(\lambda_2 - \lambda_1)$$

Donde hav es la función:

$$hav(\theta) = \sin^2(\frac{\theta}{2}) = \frac{1+\cos(\theta)}{2}$$

Siendo d la distancia entre los dos puntos, r el radio de la esfera (en este caso la tierra) y (ϕ_1, λ_1) y (ϕ_2, λ_2) las coordenadas de los puntos 1 y 2 (latitud, longitud). Para este trabajo se ha utilizado la implementación en R de la librería geosphere [44]. Sin embargo, dado que esta implementación requiere que las coordenadas sean proporcionadas como latitud y longitud y, tal y como hemos visto antes, las coordenadas de las espiras vienen dadas en formato EPSG:25830, es necesario realizar primero una conversión de estos datos. Esta conversión se ha realizado mediante la librería de R [15]. Otro problema que se puede encontrar en los datos de tráfico es el hecho de que haya filas duplicadas, lo que dificulta los cálculos, puesto que algunos puntos de medida tienen el mismo dato más de una vez. Esto se ha corregido eliminando todas las filas repetidas.

5.5 Estudio de los datos

Con todos los parámetros escogidos y los conjuntos de datos creados, se ha procedido a establecer la relación que mantienen los elementos contaminantes, con la intensidad del tráfico y los parámetros climatológicos. Para ello se ha utilizado la librería de R Openair [24], que proporciona una serie de elementos gráficos y estadísticos, altamente personalizables y especialmente diseñados para estudiar la calidad del aire.

Estación	Tráfico		Espiras en 1km	NO		NO_2		O_3		SO_2	
	ave	*sd*	*n*	*ave*	*sd*	*ave*	*sd*	*ave*	*sd*	*ave*	*sd*
Molí	13357	10302	26	10.3	19.9	28.2	20.7	46.9	25.5	2.4	2.1
Pista	34606	24462	88	23.5	33.8	45.1	27.0	47.1	25.0	3.8	3.8
Francia	20037	14277	65	8.5	18.8	26.7	23.2	50.2	25.2	2.3	2.3
Viveros	33214	24746	100	9.7	21.5	29.4	24.2	45.2	28.5	2.7	2.5
Bulevar	11352	8555	39	12.8	27.5	29.2	22.1	48.5	28.3	2.2	2.1
UPV	11987	8938	27	7.6	17.9	24.3	24.1	56.3	27.3	2.2	3.1

Tabla 5.3: Medias (*ave*) y desviación típica (*sd*) del tráfico y los contaminantes en cada una de las estaciones y número de espiras en un radio de 1km (*n*) de cada estación.

En la tabla 5.3 se puede ver un resumen de los conjuntos de datos: medias y desviaciones típicas de las seis estaciones analizadas (la media de tráfico se calcula con los sensores ubicados en un radio de 1km de la estación).

[5] format: http://www.inside-r.org/r-doc/base/format
[6] as.Date: http://www.inside-r.org/r-doc/base/as.Date

Si analizamos la intensidad del tráfico, Pista de Silla y Viveros son las estaciones con la media más alta de vehículos por hora. Pista de Silla es la estación con el mayor número de espiras de medición de tráfico cercanas, después de Viveros. En la figura 5.3 podemos ver la evolución de la intensidad horaria del tráfico asociado a la estación de la Pista de Silla, dependiendo de diferentes factores temporales. En las gráficas de distribución horaria, tanto por días (arriba) como la media de todos los días (abajo-izquierda), se pueden ver claramente tres picos de intensidad horaria en el tráfico, que coinciden con las horas de entrada/salida de empresas y centros escolares, por la mañana (7-9h.), al mediodía (12-15h.) y por las tardes (17-20h.). Los fines de semana, sin embargo, presentan solo dos picos que coinciden con la hora de comer (12-14h.) y la hora de cenar o regresar a casa (18-22h.). Como se puede observar, alrededor de la Pista de Silla se pueden ver picos de intensidad de hasta 60000 vehículos por hora. Si atendemos a la evolución del tráfico medio durante toda la semana (abajo-derecha), podemos ver como el tráfico mantiene hasta el miércoles una tendencia ascendente, disminuye el jueves y vuelve a aumentar el viernes. El sábado y domingo presentan los mínimos en la intensidad horaria para todas las estaciones. En cuanto a la evolución por meses (abajo-centro), se puede observar que el mes con más afluencia de tráfico es julio, mientras que agosto y diciembre son los meses con menos intensidad de vehículos.

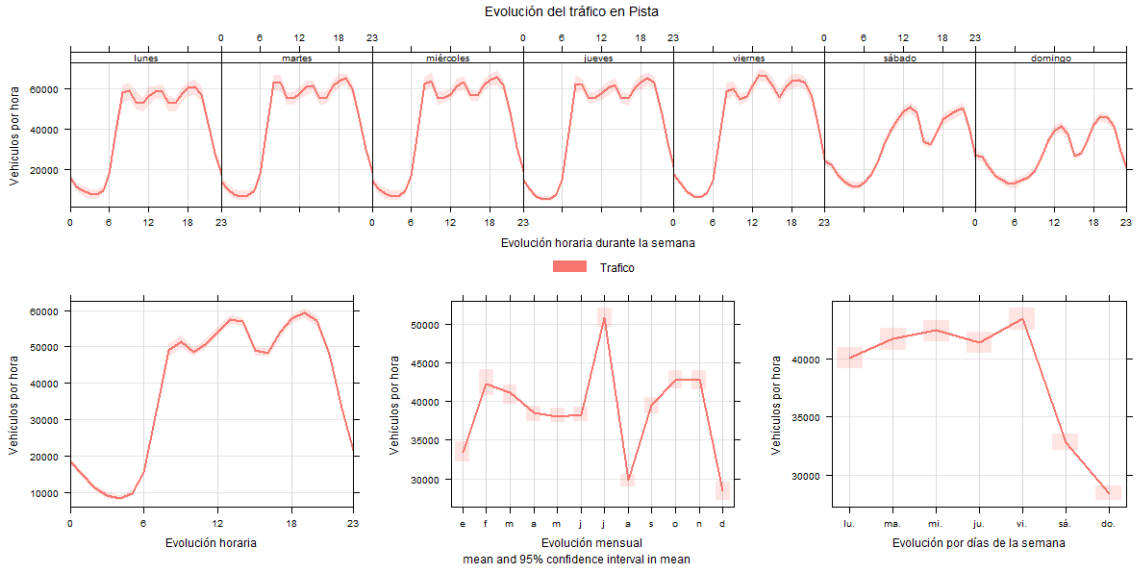

Figura 5.3: Evolución de la intensidad del tráfico relacionado con la estación de Pista de Silla dependiendo de la hora de la semana (arriba), la hora del día (abajo-izquierda), el mes (abajo-centro) y el día de la semana (abajo-derecha).

Respecto a los niveles de polución, la estación de Pista de Silla vuelve a presentar los niveles máximos para todos los parámetros, excepto O_3. El caso contrario podemos verlo en los datos de la estación de la UPV. Estos resultados pueden ir asociados a las situaciones específicas de las estaciones: Pista de Silla está ubicada en el centro de la ciudad, rodeada de calles, a menudo congestionadas y por tanto vulnerable a altos niveles de tráfico y por consiguiente, polución. Sin embargo, la estación de la UPV se encuentra dentro de la Universidad, de manera que es la que menos puntos de tráfico tiene a su alrededor.

La figura 5.4 representa la distribución de la media del nivel de los cuatro contaminantes, medidos en la estación de la Pista de Silla, dependiendo de diferentes factores temporales. Se ve una clara correspondencia entre los niveles de contaminación medidos y algunos de estos factores. Por ejemplo, se aprecia un nivel más bajo de contaminantes durante el fin de semana y los meses de verano donde, como hemos visto en la figura 5.3, el tráfico no es tan intenso. Podemos observar un pico en marzo, en el SO_2. Este es posi-

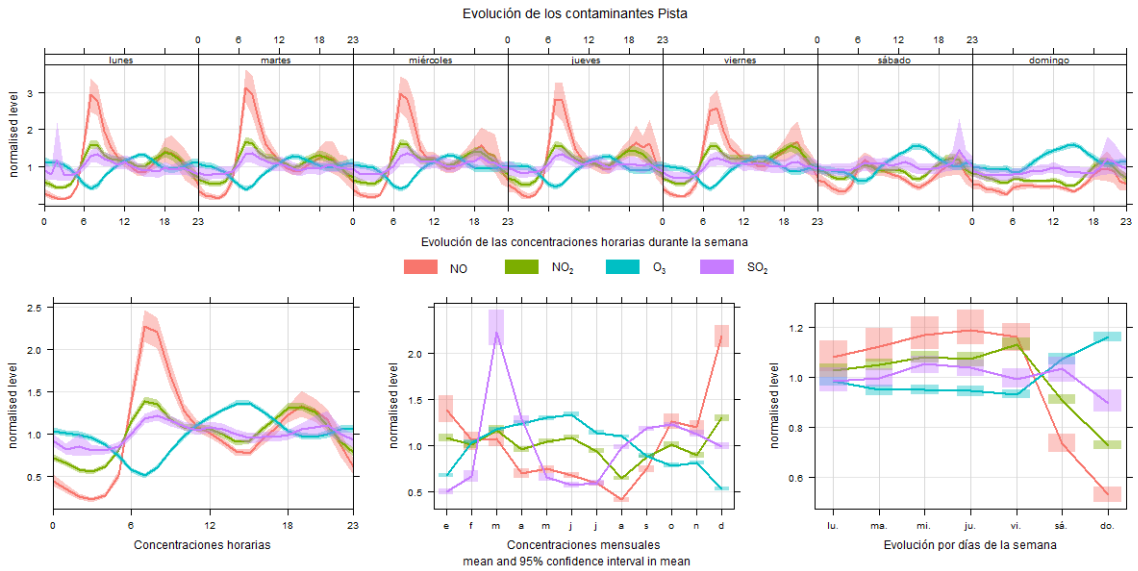

Figura 5.4: Distribución de la media de los cuatro contaminantes en la estación de Pista de Silla, dependiendo de la hora de la semana (arriba), la hora del día (abajo-izquierda), el mes (abajo-centro) y el día de la semana (abajo-derecha). Los datos están normalizados.

blemente un fenómeno local, causado por las Fallas, que concluyen el 19 de marzo con la cremà (quema de cientos de monumentos creados con materiales combustibles). También se puede ver una correlación negativa entre el Ozono (O_3) y los NOx. Esto puede ser explicado si consideramos que parte de la generación del O_3 ocurre cuando los Óxidos de nitrógeno (NO_x) reaccionan en la atmósfera en presencia de luz solar [18]. Este efecto se puede observar en los altos niveles de Ozono alrededor del mediodía y en verano. El extraño fenómeno de altos valores de O_3 en los fines de semana ha sido detectado en varias ciudades. En [43], los autores defienden que "la causa primaria del alto nivel de O_3 en fines de semana se debe a la reducción de la emisión de los Óxidos de nitrógeno (NO_x) en componentes orgánicos volátiles (VOC)".

Este tipo de correlaciones o dependencias entre los parámetros podemos verlos en las siguientes gráficas.

La figura 5.5 muestra la correlación existente entre los parámetros contaminantes y los parámetros meteorológicos para la estación de Molí del Sol. Los colores intensos muestran una fuerte correlación entre los parámetros. El signo del número (que varía entre -100 y 100) nos indica si la correlación es positiva (cuando un parámetro aumenta, el otro también aumenta) o negativa (cuando un parámetro aumenta, el otro disminuye su valor). En concreto, podemos ver una fuerte correlación positiva entre en NO y el NO_2. Esta dependencia entre los parámetros centra su lógica en el hecho de que el NO_2 deriva del NO, cuanto más NO se forme más posibilidades hay de que derive en NO_2. El ejemplo opuesto podemos verlo en la relación entre los NO_x (NO y NO_2) con el Ozono (O_3). Dado que el O_3 se forma a partir de los NO_x tiene sentido que cuando el valor de O_3 aumente, el valor de NO_x disminuya y viceversa. En la figura 5.6 podemos ver más a fondo esta correlación con los datos de la estación de la Avenida de Francia. Esta gráfica muestra la distribución en los valores de NO_2 y O_3 para los meses de verano de 2014 y en ella podemos ver claramente la correlación negativa entre estos dos parámetros. Cada punto en el eje x de la figura es la media diaria de las concentraciones, mientras que la línea uniforme es la tendencia seguida por estos datos. Esta tendencia, generada con Openair, se calcula mediante *Generalized Additive Modelling* [25] usando la librería mgcv [91].

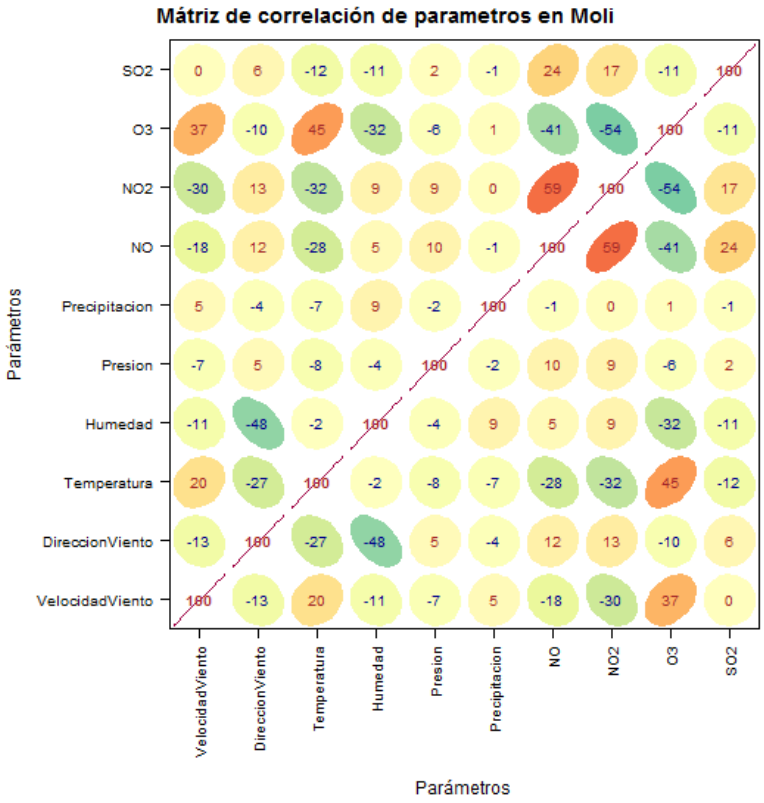

Figura 5.5: Correlación entre los parámetros contaminantes y los meteorológicos.

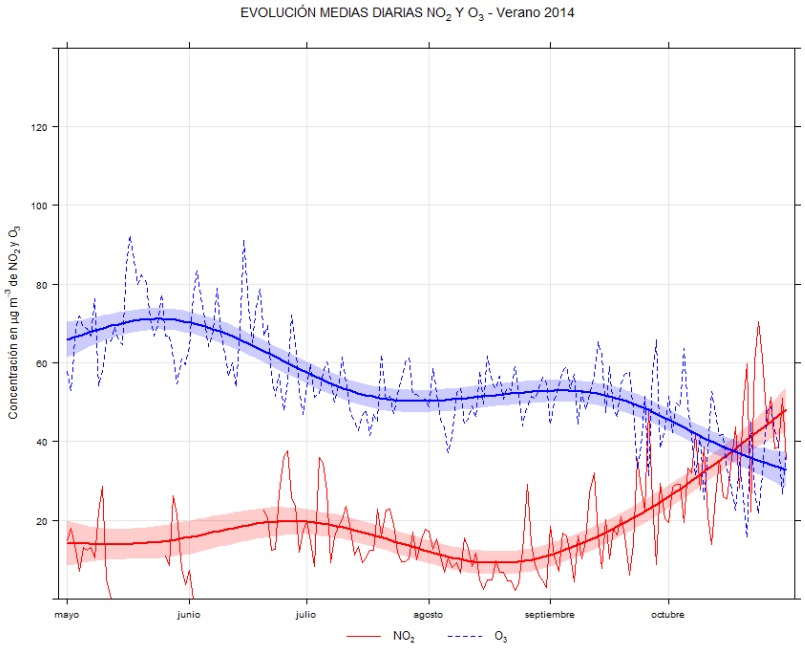

Figura 5.6: Concentración de NO_2 y O_3 en la estación de la Avenida de Francia para el verano de 2014.

La figura 5.5 nos sirve también para ver la manera en que los parámetros meteorológicos afectan a los contaminantes. Por ejemplo, si observamos el punto de unión entre el O_3 y la velocidad del viento o la temperatura, podemos ver como el aumento de cualquiera de estos dos parámetros hace que el Ozono aumente también, sin embargo la humedad hace que disminuya. Por contra, la velocidad del viento y la temperatura ejercen una in-

fluencia inversa con los NO_x haciendo que los contaminantes se dispersen y disminuyan su concentración cuando el valor de estos parámetros meteorológicos aumenta.

Es importante, viendo estas relaciones, estudiar mejor los parámetros y establecer aquellas relaciones que resulten útiles de cara a realizar con ellos la predicción de los contaminantes. La figura 5.7 representa la Rosa de los vientos de Valencia. Este tipo de representaciones gráficas muestran cuáles son las velocidades y direcciones del viento más habituales en una zona. Los sectores de las rosas de los vientos muestran el origen del viento, por ejemplo, el viento del norte se muestra en el ángulo 0º o 360º (la parte superior), mientras que los colores denotan la intensidad del viento. En el caso de Valencia el viento que cobra más protagonismo por frecuencia es el poniente, mientras que el que llega con más fuerza es el viento del Noreste.

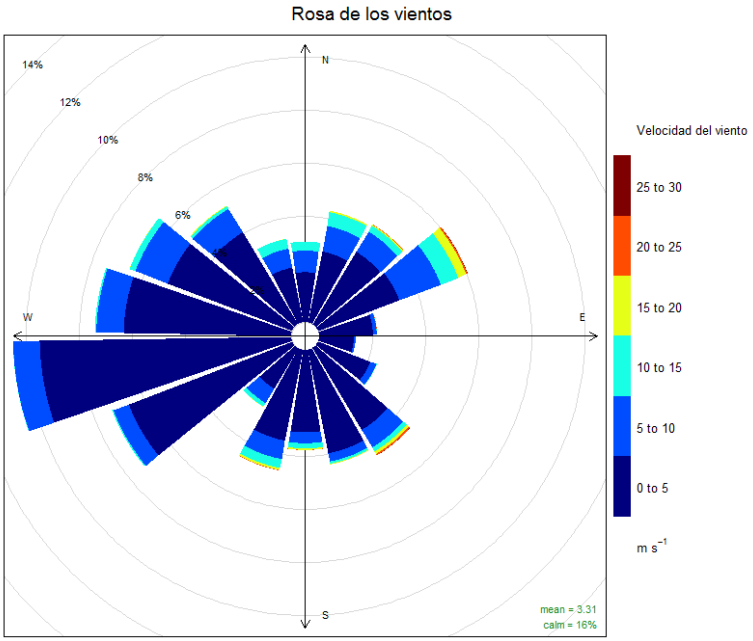

Figura 5.7: Rosa de los vientos de Valencia.

Si aplicamos este mismo tipo de gráfica a los contaminantes, obtenemos una gráfica polar de frecuencias, donde podemos ver cómo evolucionan los parámetros contaminantes en función de la dirección y velocidad del viento [37]. En estos gráficos se clasifican las mediciones por intervalos de dirección, cuyos valores medios son los 16 rumbos (NNE, NE, ENE, E,...), y por intervalos de velocidad. Estas gráficas reflejan la influencia de la dirección del viento en la concentración o dispersión de los contaminantes [60]. La figura 5.8 muestra este tipo de gráficas. En ella podemos ver la dispersión del NO_2 en la estación de Bulevar Sur dependiendo de la velocidad y dirección del viento. En este caso, a mayor velocidad del viento, menos concentración de NO_2.

La figura 5.9 muestra una rosa de polución, donde la velocidad del viento es sustituida por la concentración de un contaminante. En este tipo de gráfica podemos ver la relación existente entre dos contaminantes y la dirección del viento. En esta figura se ha dividido la concentración de NO en 4 niveles y para cada uno de ellos se ha creado una rosa de polución con los niveles de NO_2, dependiendo de la dirección del viento. Podemos ver como los niveles de NO_2 aumentan (color más intenso) conforme aumentan los niveles de NO, como hemos podido comprobar también en la matriz de correlaciones de la figura 5.5.

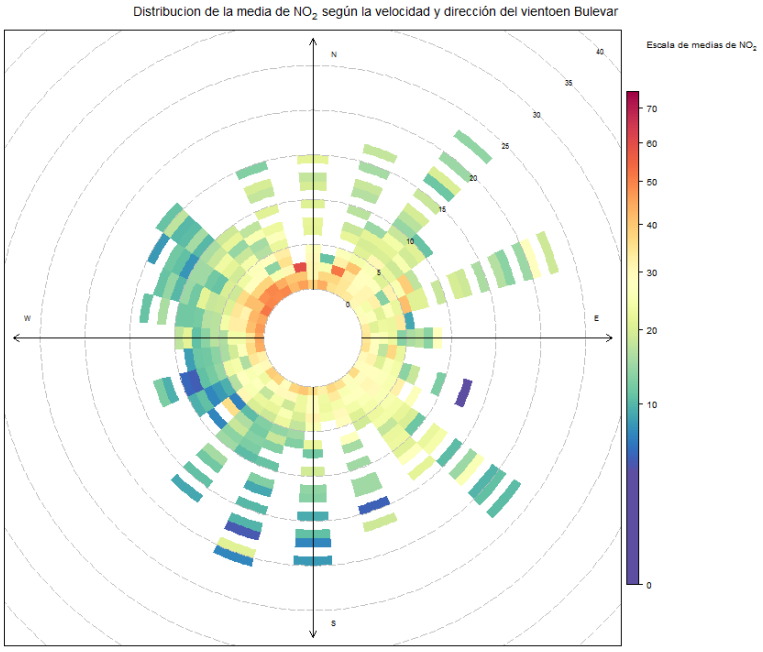

Figura 5.8: Distribución de la media de NO_2 dependiendo de la velocidad y dirección del viento en la estación de Bulevar Sur.

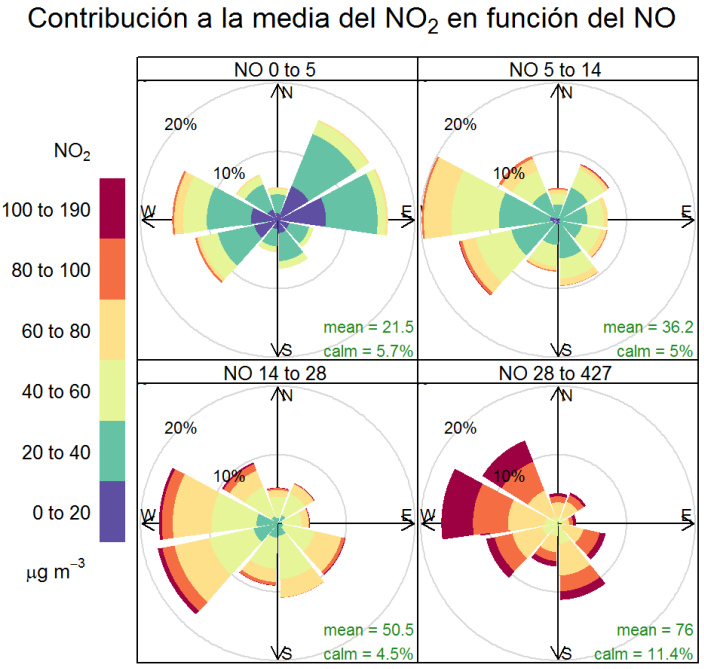

Figura 5.9: Relación entre los niveles de NO y NO_2 en la estación de la Pista de Silla.

La figura 5.10 nos muestra una gráfica polar de la distribución del NO_2 en la estación de Viveros en función del viento, tal y como hemos visto en las gráficas anteriores. Sin embargo, este tipo de representación responde a un modelizado matemático que suaviza la cuadrícula y proporciona una superficie continua. La utilidad de esta gráfica es la de identificar fuentes potenciales de origen [37]. Como vemos en esta figura, el contaminante proviene sobre todo del Suroeste. Esta conclusión cobra sentido si tenemos en cuenta los datos proporcionados anteriormente sobre la ubicación de las estaciones (figura 5.1), dado que la estación de Viveros únicamente tiene puntos de tráfico cercanos al Oeste de los jardines.

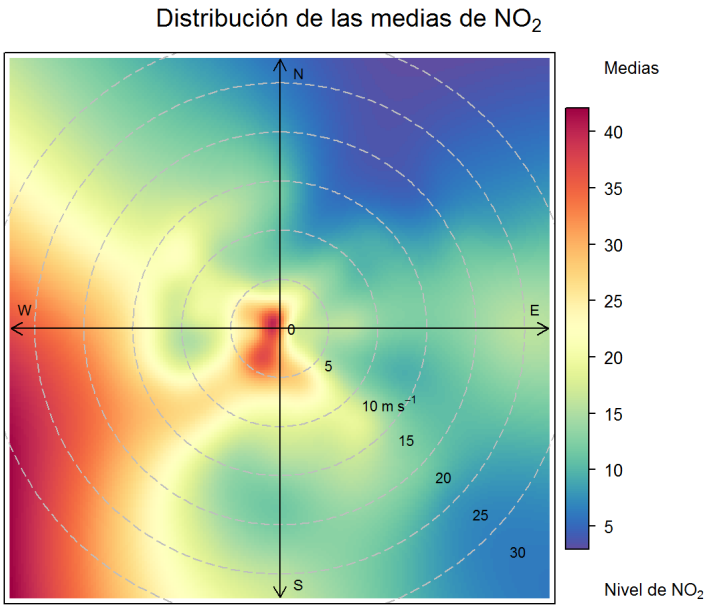

Figura 5.10: Gráfica polar de la media de NO_2 en la estación de Viveros.

Por último, sabemos que la mayoría de los contaminantes seleccionados provienen, total o parcialmente, de las emisiones originadas por la quema de combustibles en los vehículos. En la figura 5.11 podemos observar la relación entre el NO_2, la intensidad del tráfico y la velocidad del viento. Se puede deducir, en las gráficas de medias horarias (arriba y abajo-izquierda), como el contaminante sigue la tendencia del tráfico (fuente de emisión) y aumenta a lo largo de la mañana, hasta que llegado un momento (sobre las horas del mediodía), aumenta la velocidad del viento y dispersa el NO_2, haciendo que su valor disminuya. En la gráfica semanal (abajo-derecha) se puede ver más claramente está correlación positiva entre el contaminante y el tráfico.

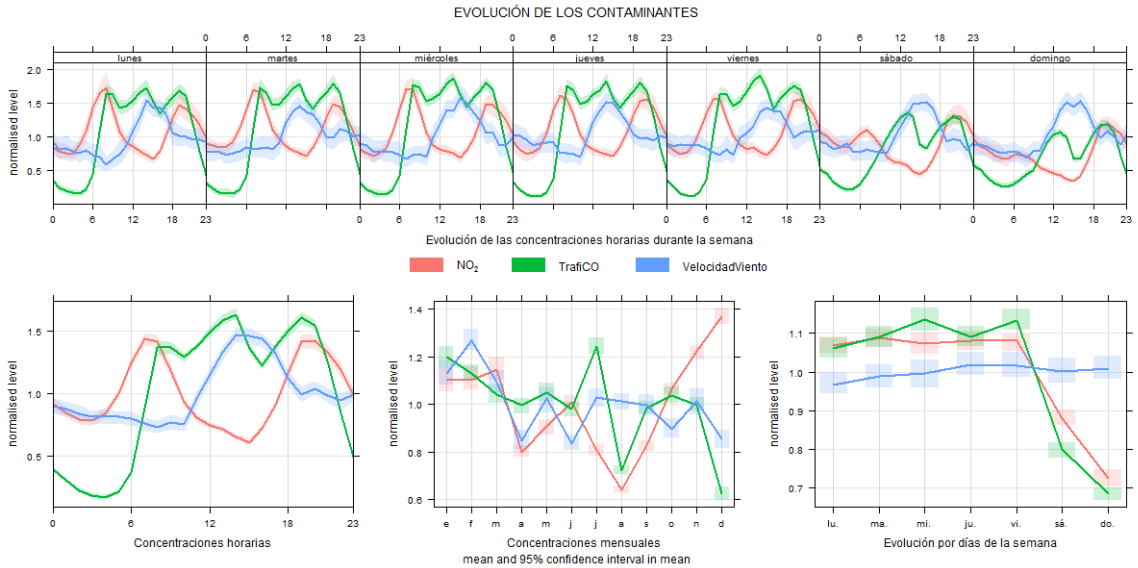

Figura 5.11: Evolución de las medias de NO_2, Intensidad horaria del tráfico y Velocidad del Viento. Los datos están normalizados.

El análisis de datos estudiado en esta sección, pone de manifiesto el hecho de que los parámetros contaminantes se ven especialmente afectados por tres factores críticos: La velocidad del viento, la temperatura y el número de vehículos por hora. Es obvio que la concentración y dispersión de los contaminantes depende directa o indirectamente de estos parámetros. Estos hechos avalan el uso de estos datos como base para crear el modelo de predicción de los contaminantes durante el resto de este trabajo.

De esta manera, hemos construido conjuntos de datos que sirven de base para predecir las concentraciones de contaminantes a partir de la intensidad del tráfico y los parámetros meteorológicos y además, después de ver las relaciones entre los parámetros, se ha añadido a los modelos, el nivel de los contaminantes, de la temperatura y del tráfico 3 horas antes (periodo de tiempo en el que se pueden obtener los datos actuales más próximos a la predicción) y adicionalmente: el nivel de NO y SO_2 3 horas antes, en el caso del NO_2; el nivel del NO_2 y el SO_2 3 horas antes, en el caso del NO y el nivel del NO y NO_2 3 horas antes, en el caso del O_3.

CAPÍTULO 6
Predicción de los niveles de contaminación

En esta sección estudiaremos diferentes técnicas de regresión con el fin de obtener la predicción de los cuatro contaminantes en las seis estaciones de Valencia. Con este fin, realizaremos diferentes experimentos utilizando los datos históricos y comparándolos mediante diferentes medidas de rendimiento, determinando cuál de las técnicas de regresión es la más óptima para realizar la predicción.

6.1 Introducción

Como se ha visto en el capítulo 5, se han creado en total seis conjuntos de datos. Cada uno de estos *datasets* contiene los datos reales de los cuatro contaminantes, la intensidad del tráfico relacionado con la estación y los cinco parámetros meteorológicos desde enero de 2013 hasta diciembre de 2015. El objetivo ahora es, para cada contaminante en cada estación, determinar la técnica de regresión que mejor funcione para predecir el valor de dicho contaminante, es decir, aquella que obtenga el menor error en los resultados y por tanto, su valor predicho se aproxime más al valor real.

Como hemos descrito anteriormente, las técnicas de regresión son esencialmente el entrenamiento o aprendizaje de funciones de valores reales a partir de datos. Esto se hace normalmente eligiendo una clase de funciones y construyendo una función que minimice la diferencia entre los valores predichos y los reales [36].

6.2 Experimentos

Para poder determinar el mejor método de predicción para cada contaminante, se han ejecutado experimentos con las diferentes técnicas de regresión explicadas en el capítulo 3 y se han comparado los resultados obtenidos por cada una de ellas. Cada una de estas técnicas de regresión ha sido utilizada en su implementación en R [74] con los parámetros por defecto, a no ser que se especifiquen otros:

- *Linear Regression* (*lr*) [64]
- *Quantile regression* (*qr*) [56] con método *lasso*
- *K nearest neighbours* (*IBKreg*) con $k = 10$ [64]
- Un árbol de decisión para regresión (*M5P*) [64]

- *Random Forest* (*RF*) [59]

Para comparar los resultados de los modelos, se han introducido tres modelos que servirán de referencia (*baseline models*):

- Un modelo que siempre calcula la media del conjunto de entrenamiento (*TrainMean*)
- Un modelo que siempre calcula la media del conjunto de prueba (*TestMean*)
- Un modelo que siempre calcula el valor del contaminante 3 horas antes (*X3H*).

Para evaluar los resultados se ha usado *Root Mean Squared Error* (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(C_m^i - C_0^i)^2}{N}}$$

Donde C_m^i es la predicción (concentración pronosticada) para la hora i y C_0^i es la concentración real observada para la hora i. Este valor nos proporciona el valor de error cuadrático cometido por nuestro modelo.

6.2.1. *Data Splitting*

Para realizar una comparación entre los métodos de manera fiable se ha utilizado la técnica de *Data Splititng* para series temporales. En concreto, en nuestro caso hemos utilizado un muestreo de conveniencia (*convenience sampling*) [76], donde el conjunto de datos se divide en bloques discretos, por ejemplo, intervalos de tiempo. Para nuestros experimentos el conjunto se ha dividido en meses y se han realizado 4 ejecuciones por cada contaminante de cada estación, utilizando siempre los últimos meses de 2015 como *test* y el resto como *train*, del siguiente modo:

- Primera ejecución: 1 mes de *test*, resto de meses (35) de *train*.
- Segunda ejecución: 2 meses de *test*, resto de meses (34) de *train*.
- Tercera ejecución: 3 meses de *test*, resto de meses (33) de *train*.
- Cuarta ejecución: 4 meses de *test*, resto de meses (32) de *train*.

Al final de la cuarta ejecución, el resultado final será la media del RMSE de las cuatro pasadas. En la Figura 6.1 se puede ver un esquema gráfico de las ejecuciones realizadas.

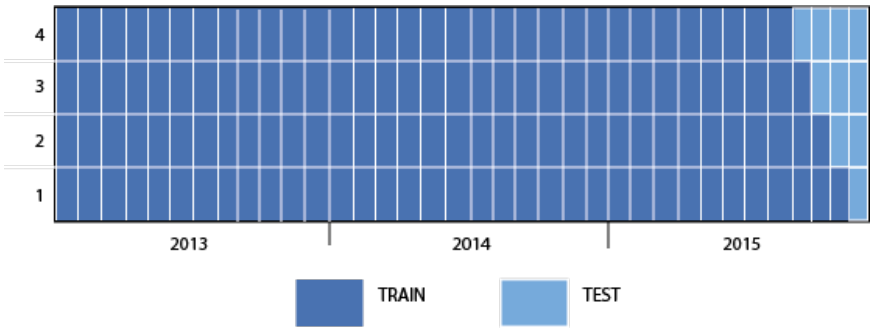

Figura 6.1: Intervalos temporales para los experimentos mediante *Data Splitting*

6.3 Resultados

	TrainMean	*TestMean*	*X3h*	*lr*	*qr*	*IBKreg*	*M5P*	*RF*
NO	56.79	49.93	58.85	47.38	53.79	46.95	39.76	**36.88**
NO_2	27.33	21.53	22.17	19.12	19.71	19.82	15.79	**14.35**
O_3	33.99	22.24	19.74	19.11	18.67	16.95	13.02	**11.64**
SO_2	1.97	1.52	1.6	1.4	1.6	1.62	1.45	**1.38**

Tabla 6.1: Resultados en RMSE de los diferentes modelos de regresión para la estación de Bulevar Sur. La mejor predicción está resaltada en negrita.

	TrainMean	*TestMean*	*X3h*	*lr*	*qr*	*IBKreg*	*M5P*	*RF*
NO	58.09	53.01	57.68	51.43	54.24	50.82	48.81	**44.92**
NO_2	39.66	29.49	28.54	26.78	27.94	28.95	21.68	**21.52**
O_3	30.06	22.7	17.48	17.44	17.23	17.46	14.05	**12.06**
SO_2	3.25	2.88	2.82	2.47	2.82	2.88	**2.22**	2.26

Tabla 6.2: Resultados en RMSE de los diferentes modelos de regresión para la estación de Avenida de Francia. La mejor predicción está resaltada en negrita.

	TrainMean	*TestMean*	*X3h*	*lr*	*qr*	*IBKreg*	*M5P*	*RF*
NO	40.89	37.68	44.41	34.81	39.42	33.26	30.28	**28.68**
NO_2	33.73	28.73	26.52	23.27	25.00	27.32	20.79	**19.76**
O_3	29.68	24.70	19.63	19.42	18.61	18.81	13.65	**11.98**
SO_2	1.66	1.41	1.85	1.59	1.85	1.47	1.49	**1.33**

Tabla 6.3: Resultados en RMSE de los diferentes modelos de regresión para la estación de Molí del Sol. La mejor predicción está resaltada en negrita.

	TrainMean	*TestMean*	*X3h*	*lr*	*qr*	*IBKreg*	*M5P*	*RF*
NO	58.79	53.15	61.99	47.31	53.01	47.27	40.43	**37.1**
NO_2	24.93	23.35	22.87	16.94	17.29	18.86	18.01	**14.02**
O_3	34.34	24.05	20.34	18.48	18.16	19.43	15.21	**13.75**
SO_2	1.62	1.61	1.78	1.52	1.67	1.97	**1.48**	1.52

Tabla 6.4: Resultados en RMSE de los diferentes modelos de regresión para la estación de Pista de Silla. La mejor predicción está resaltada en negrita.

	TrainMean	*TestMean*	*X3h*	*lr*	*qr*	*IBKreg*	*M5P*	*RF*
NO	62.41	57.86	69.51	55.46	59.88	53.22	48.8	**47.36**
NO_2	47.54	37.22	38.98	32.89	35.36	34.25	27.28	**25.36**
O_3	37.17	25.41	21.91	22.89	22.27	20.53	15.57	**13.74**
SO_2	1.62	1.37	1.72	1.39	1.71	1.57	1.61	**1.31**

Tabla 6.5: Resultados en RMSE de los diferentes modelos de regresión para la estación de UPV. La mejor predicción está resaltada en negrita.

	TrainMean	*TestMean*	*X3h*	*lr*	*qr*	*IBKreg*	*M5P*	*RF*
NO	35.65	34.12	40.42	32.2	35.26	30.69	31.13	**26.12**
NO_2	23.62	22.13	20.96	18.42	18.98	20.56	15.89	**14.14**
O_2	35.74	32.33	24.91	23.62	22.79	24.26	18.04	**16.53**
SO_2	1.12	1.01	1.22	1.03	1.22	1.23	**1.02**	1.04

Tabla 6.6: Resultados en RMSE de los diferentes modelos de regresión para la estación de Viveros. La mejor predicción está resaltada en negrita.

6.4 Discusión

Las tablas anteriores (Tabla 6.1, Tabla 6.3, Tabla 6.2, Tabla 6.4, Tabla 6.5 y Tabla 6.6) contienen el RMSE de los modelos de regresión para las predicciones de los cuatro contaminantes en cada una de las estaciones. Observando estos resultados, podemos concluir que los modelos de *Machine Learning* utilizados son capaces de mejorar los resultados de los modelos básicos en todos los casos. El menor error se obtiene aplicando *Random Forest* para todos los contaminantes en todas las estaciones, salvo en el caso del dióxido de azufre (SO_2) en Pista de Silla, Viveros y Avenida de Francia. Pese a este último dato y para simplificar el problema, *Random Forest* será el modelo de predicción utilizado en los demás experimentos de este trabajo.

6.5 Modelos considerando la dirección del viento

Los modelos que hemos creado en esta sección consideran la velocidad del viento como parámetro, pero no la dirección. Teniendo en cuenta que este parámetro puede contribuir significativamente a la dispersión de los contaminantes, como se ha discutido en el capítulo 2, se han probado diferentes técnicas para incorporarlo a los modelos.

Una manera simple de usar la dirección del viento como parámetro, es considerar como par de atributos el seno y coseno del ángulo. En nuestro caso, este método ha obtenido pésimos resultados empeorando los errores obtenidos anteriormente, de modo que se ha decidido abordar el problema desde un enfoque diferente. Teniendo en cuenta que el tráfico está generando la mayor parte de la contaminación que intentamos predecir y que estos contaminantes son dispersados por el viento, se ha decidido modificar la forma en la que se calcula el tráfico alrededor de cada estación en función de la dirección del viento. Con este propósito se han probado tres maneras diferentes de seleccionar las espiras electromagnéticas que serán tenidas en cuenta en los cálculos de los niveles de intensidad del tráfico: la versión *nd* (versión original), en la que todas las espiras en un radio de 1km son tenidas en cuenta; la versión *dir* en la que solo se tienen en cuenta aquellas espiras que se encuentren a barlovento[1]; y la versión *wdir*, que tendrá en cuenta todas las espiras en un radio de 1km, pero dando mayor peso a aquellas a barlovento.

Para calcular la dirección entre las estaciones y las espiras de tráfico se ha utilizado la función *bearing rhumb*, que calcula la dirección de desplazamiento o rumbo verdadero a lo largo de una línea de rumbo (loxodromia[2]) entre dos puntos. Para este trabajo se ha utilizado la implementación en R de la librería [44]. Esta función devuelve un ángulo entre 0º y 359º, aumentando en sentido contrario a las agujas del reloj (Este, Norte, Oeste, Sur), siendo 0º el Este. Sin embargo, la dirección del viento proporcionada por las

[1] El término barlovento hace referencia a la dirección de donde viene el viento, con respecto a un punto o lugar determinado [73].

[2] Se denomina loxodrómica o loxodromia a la línea que une dos puntos cualesquiera de la superficie terrestre cortando a todos los meridianos con el mismo ángulo.

estaciones meteorológicas varían de 0º a 359º, en el sentido de las agujas del reloj (Norte, Este, Sur, Oeste) y siendo 0º el Norte. Por tanto, es necesario realizar una conversión de los ángulos recibidos en los datos climatológicos para poder compararlos con la dirección de las estaciones a las espiras. Esta conversión la realizamos del siguiente modo:

$$dv = \begin{cases} 90 - D & \text{si } 0 \leq D \leq 90 \\ 360 - (D - 90) & \text{si } 90 < D \leq 359 \end{cases}$$

Donde D es la dirección del viento en los datos meteorológicos y dv la dirección del viento calculada según el rumbo.

Los puntos de tráfico utilizados para los métodos *dir* y *wdir*, se seleccionan si se encuentran dentro de un sector de 30º a barlovento (± 30º de la dirección del viento calculada). Para seleccionar aquellos puntos que se encuentran dentro de este sector, calculamos un rango de ángulos válidos para posteriormente, seleccionar los puntos que se encuentran dentro de esta zona según la dirección a la que se encuentren de la estación. Este rango de ángulos válidos lo calculamos de la siguiente manera:

$$Rango = \begin{cases} [0, dv + 30] \cup [360 + (dv - 30), 359] & \text{si } 0 \leq dv \leq 30 \\ [0, 30 - (360 - dv)] \cup [dv - 30, 359] & \text{si } 330 \leq dv \leq 359 \\ [dv - 30, dv + 30] & \text{si } 30 < dv \leq 330 \end{cases}$$

Donde dv es el ángulo calculado anteriormente.

A continuación se detallan gráficamente los tres métodos:

- Método *nd* (Figura 6.2): Versión original, en la que el tráfico asignado a cada estación se calcula como la suma de todo el tráfico medido por las espiras que se encuentren en un radio de 1km alrededor de la estación. En este método no se tiene en cuenta la dirección del viento para ninguno de los cálculos.

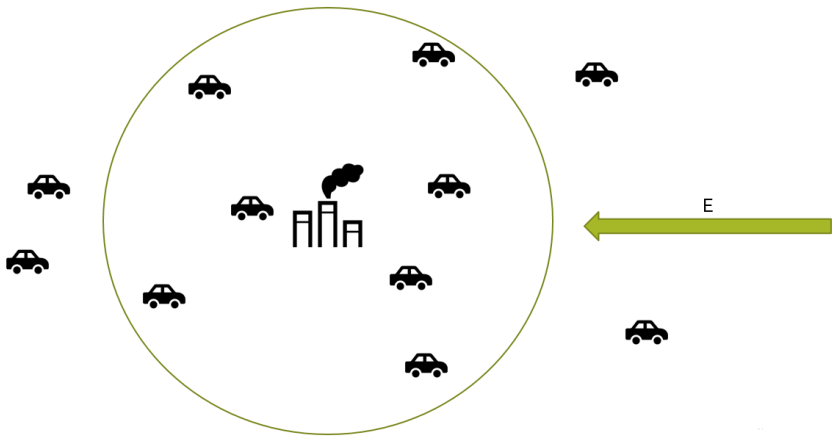

Figura 6.2: Esquema gráfico del modelo de predicción *nd*, donde se tienen en cuenta todos los puntos de tráfico en un radio de 1km alrededor de la estación de medición de contaminación. La flecha indica la dirección del viento.

- Método *dir* (Figura 6.3): En esta versión se calcula el tráfico generado en un radio de 1km alrededor de la estación, pero solo considerando aquellos sensores en un ángulo de 30º, es decir, el tráfico cuya polución generada es potencialmente arrastrada por el viento hacía la estación.

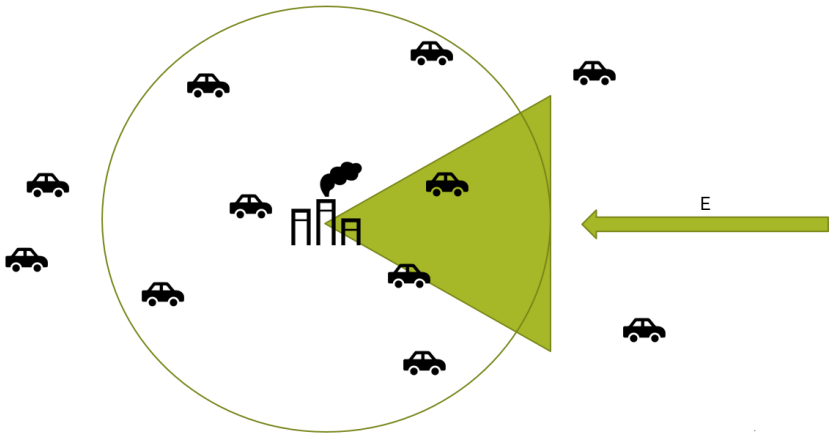

Figura 6.3: Esquema gráfico del modelo de predicción *dir*, donde se tienen en cuenta solo los puntos de tráfico en un radio de 1km alrededor de la estación de medición de contaminación que se encuentren a barlovento. La flecha indica la dirección del viento.

- Método *wdir* (Figura 6.4): Este método es una combinación del anterior método (*dir*) y el método original (*nd*). En esta versión consideramos todo el tráfico en un radio de 1km de la estación, pero le damos mayor peso a aquellos puntos de tráfico a barlovento, en un ángulo de 30º.

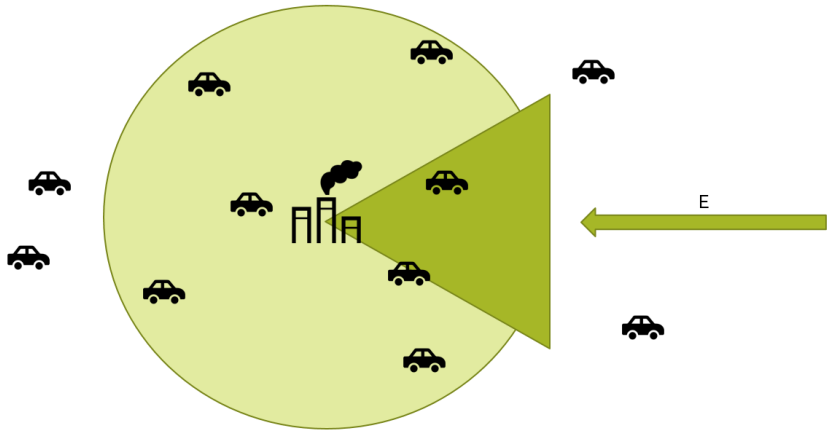

Figura 6.4: Esquema gráfico del modelo de predicción *wdir*, donde se tienen en cuenta todos los puntos de tráfico en un radio de 1km alrededor de la estación de medición de contaminación, pero dando mayor peso a aquellos que se encuentren a barlovento. La flecha indica la dirección del viento.

Para calcular el tráfico total ponderado, se clasifican los puntos según se encuentren dentro o fuera del rango calculado y se les aplica el peso correspondiente a su situación, del siguiente modo:

$$TraficoTotal = \sum_{i=1}^{n} x_i \frac{\alpha}{N} + \sum_{j=1}^{p} y_j \frac{\beta}{N}$$

Donde $x_1, x_2, \ldots, x_n$ son los valores de tráfico de los puntos dentro del rango en un radio de 1km, $y_1, y_2, \ldots, y_p$ son los valores de tráfico de los puntos fuera del rango en un radio de 1km, $N = n + p$, $\alpha = 1{,}5$ y $\beta = (N - (\alpha * n))/(N - n)$.

6.5.1. Resultados

Estación	NO			NO_2			O_3			SO_2		
	nd	*dir*	*wdir*	*nd*	*dir*	*wdir*	*nd*	*dir*	*wdir*	*nd*	*dir*	*wdir*
Bulevar	**39.59**	40.02	40.27	14.12	**13.31**	14.05	10.92	**10.64**	10.89	1.56	**1.54**	1.56
Francia	49.86	50.92	**49.81**	**21.44**	21.74	21.53	11.79	12.11	**11.70**	**2.51**	2.58	2.52
Moli	**30.69**	32.24	30.76	19.82	19.83	**19.79**	11.48	11.48	**11.44**	1.41	1.45	**1.40**
Pista	39.69	40.65	**39.40**	13.52	13.58	**13.43**	12.09	12.25	**12.06**	**1.54**	1.55	1.56
UPV	54.41	57.39	**54.33**	**26.96**	27.61	26.99	13.28	**13.17**	13.20	**1.40**	1.44	1.41
Viveros	30.37	31.23	**30.30**	15.11	15.25	**15.01**	15.62	**15.44**	15.56	1.13	1.12	**1.11**

Tabla 6.7: RSME de las predicciones con *Random Forest* para los cuatro contaminantes dependiendo del método utilizado para considerar la dirección del viento: *nd* (la dirección del viento no se tiene en cuenta), *dir* (solo se tienen en cuenta los sensores de tráfico a barlovento), *wdir* (se le da más peso a los sensores de tráfico a barlovento). La mejor predicción está resaltada en negrita.

6.5.2. Discusión

La dirección del viento puede contribuir significativamente en la dispersión de los contaminantes, en este caso aquellos originados por emisiones del tráfico. En la Tabla 6.7 comparamos los resultados en RMSE de las seis estaciones usando tres escenarios diferentes e incluyendo los métodos que consideran la dirección del viento para calcular el nivel de tráfico que afecta a cada estación: *nd* (la dirección del viento no se tiene en cuenta), *dir* (solo se tienen en cuenta los sensores de tráfico a barlovento), *wdir* (se le da más peso a los sensores de tráfico a barlovento).

Los resultados muestran que el rendimiento de los métodos depende drásticamente de los contaminantes y las estaciones. Los métodos *dir* y especialmente *wdir*, son capaces de mejorar la predicción en partículas relacionadas directamente con el tráfico, como NO y el Ozono, probablemente porque estás técnicas de modelado con la dirección del viento se basan en la selección de los puntos de emisión de tráfico. En este caso, para SO_2, se puede observar que estos métodos no consiguen superar los resultados del método original en todos los casos, posiblemente porque el tráfico no tiene tanta influencia en su origen.

En cualquier caso, se puede observar variación en los resultados dependiendo de la estación y el contaminante, cuya razón probablemente se deba a las diferentes localizaciones y situaciones de las estaciones. Estos resultados habría que estudiarlos más a fondo con datos que no se han tenido en cuenta como la topografía y demografía de la zona: edificios, zonas verdes, población, industrias, etc.

6.6 Evaluación del modelo

Para evaluar el modelo vamos a comparar los datos reales observados y las predicciones que nuestro modelo ha realizado. Con este propósito usaremos las predicciones generadas con *Random Forest* para cada estación y contaminante, utilizando el método *wdir* para calcular el tráfico dependiendo de la dirección del viento y usando como test los últimos cuatro meses[3] de 2015.

En la Tabla 6.8 podemos ver el sesgo medio (MB) en los valores predichos para cada contaminante en cada una de las estaciones, cuya fórmula es:

[3]Se utilizan los últimos cuatro meses con datos reales, es decir, las horas en las que no haya dato real se saltan y se sustituyen por la hora anterior que contenga dato, hasta completar cuatro meses de datos horarios.

$$MB = \sum_{i=1}^{N} \frac{(C_m^i - C_0^i)}{N}$$

Donde C_m^i es la concentración predicha para la hora i y C_0^i es la concentración observada para la hora i. Este valor nos proporciona información sobre la tendencia del modelo a subestimar o sobreestimar el valor de un parámetro sobre el valor real.

Por otra parte, en la misma tabla podemos ver el coeficiente de correlación de Pearson (r), que mide la relación lineal existente entre dos variables, dando una idea del grado de relación existente entre ambas, de forma que $r = 1$ implica que existe relación directa perfecta, y $r = 0$ que no se puede comprobar que exista relación lineal entre ambas. El coeficiente de correlación se puede calcular del siguiente modo:

$$r = \frac{\sum_{i=1}^{N}(C_m^i - \bar{C}_m)(C_0^i - \bar{C}_0)}{\sqrt{\sum_{i=1}^{N}(C_m^i - \bar{C}_m)}\sqrt{\sum_{i=1}^{N}(C_0^i - \bar{C}_0)}}$$

Donde C_m^i es la concentración predicha para la hora i y C_0^i es la concentración observada para la hora i, $\bar{C}_m$ es el valor promedio modelado durante el periodo N y $\bar{C}_0$ el valor promedio observado durante el periodo N.

Estación	NO		NO_2		O_3		SO_2	
	MB	*r*	*MB*	*r*	*MB*	*r*	*MB*	*r*
Bulevar	-5.66	0.77	-5.60	0.82	4.18	0.91	-0.07	0.51
Francia	-5.97	0.64	-6.60	0.83	0.85	0.89	-0.22	0.76
Moli	-2.58	0.70	-5.15	0.81	0.91	0.89	0.60	0.30
Pista	-0.85	0.76	-2.40	0.81	0.67	0.86	0.98	0.48
UPV	-3.81	0.70	-4.92	0.83	3.08	0.90	0.27	0.45
Viveros	-0.44	0.69	0.21	0.81	-0.48	0.89	0.32	0.40

Tabla 6.8: Sesgo medio (MB) y coeficiente de correlación (r) por contaminante y estación para la versión *wdir* usando como test los últimos cuatro meses de 2015.

La Tabla 6.8 nos muestra que nuestro modelo de predicción, en general, tiende a subestimar los valores de NO y NO_2 y a sobreestimar los de Ozono y SO_2. En cuanto al coeficiente de correlación, podemos ver que el promedio de valores del O_3 es el que más relación tiene con el promedio observado realmente en las estaciones, mientras que el SO_2 es el que menos relación lineal tiene con su valor observado. En la Figura 6.5 podemos ver una gráfica comparativa entre los datos reales (en rojo) y los datos predichos por el modelo (en azul) para el Ozono en la estación de Viveros.

En [32] se evalúa el sistema CALIOPE de pronóstico de la calidad del aire siguiendo las métricas usadas en esta memoria. Para este propósito, proponen unos criterios de calidad en función del contaminante y clasifican el modelo según 5 clases:

- Muy Bueno (MB)
- Bueno (B)
- Aceptable (A)
- Malo (M)
- Muy Malo (MM)

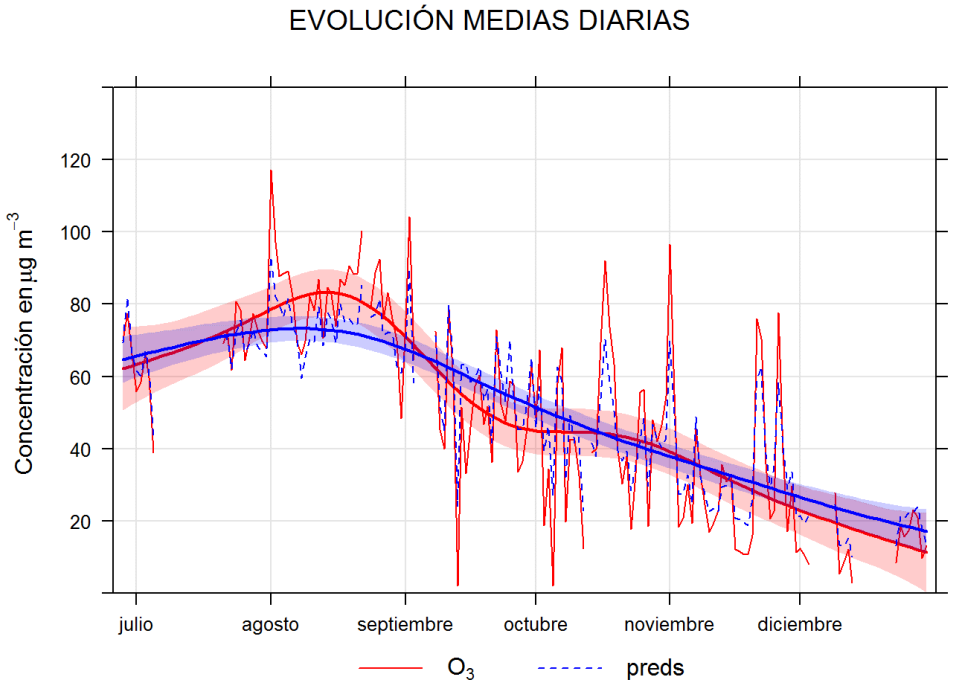

Figura 6.5: Comparación entre los datos reales y los predichos por el modelo para el O_3 en la estación de Viveros.

Los valores que deben tenerse en cuenta para clasificar el modelo en función de estos criterios son los siguientes:

Contaminante	*Clasificación*	*BIAS* ($\mu g/m^3$) $MB = 0$*	$r = 1$*	*RMSE* ($\mu g/m^3$) $RMSE = 0$*
O_3	Muy Bueno (MB)	$\|MB\| < 5$	$r > 0{,}70$	$RMSE < 10$
	Bueno (B)	$5 \leq \|MB\| < 10$	$0{,}50 \leq r < 0{,}70$	$10 \leq RMSE < 20$
	Aceptable (A)	$10 \leq \|MB\| < 20$	$0{,}30 \leq r < 0{,}50$	$20 \leq RMSE < 30$
	Malo (M)	$20 \leq \|MB\| < 30$	$0{,}10 \leq r < 0{,}30$	$30 \leq RMSE < 40$
	Muy Malo (MM)	$\|MB\| \geq 30$	$r < 0{,}10$	$RMSE \geq 30$
NO_2	Muy Bueno (MB)	$\|MB\| < 5$	$r > 0{,}60$	$RMSE < 5$
	Bueno (B)	$5 \leq \|MB\| < 10$	$0{,}40 \leq r < 0{,}60$	$5 \leq RMSE < 15$
	Aceptable (A)	$10 \leq \|MB\| < 20$	$0{,}20 \leq r < 0{,}60$	$15 \leq RMSE < 25$
	Malo (M)	$20 \leq \|MB\| < 30$	$0{,}10 \leq r < 0{,}20$	$25 \leq RMSE < 35$
	Muy Malo (MM)	$\|MB\| \geq 30$	$r < 0{,}10$	$RMSE \geq 35$
SO_2	Muy Bueno (MB)	$\|MB\| < 3$	$r > 0{,}40$	$RMSE < 5$
	Bueno (B)	$3 \leq \|MB\| < 5$	$0{,}30 \leq r < 0{,}40$	$5 \leq RMSE < 15$
	Aceptable (A)	$5 \leq \|MB\| < 10$	$0{,}20 \leq r < 0{,}30$	$15 \leq RMSE < 25$
	Malo (M)	$10 \leq \|MB\| < 20$	$0{,}10 \leq r < 0{,}20$	$25 \leq RMSE < 35$
	Muy Malo (MM)	$\|MB\| \geq 20$	$r < 0{,}10$	$RMSE \geq 35$

Tabla 6.9: Criterios de clasificación del modelo según le evaluación del sistema CALIOPE. (*) Valores óptimos.

CALIOPE no evalúa los niveles de NO, que sí se miden en este trabajo, pero si los otros tres contaminantes (O_3, NO_2 y SO_2) como se puede observar en la tabla 6.9. Si tenemos en cuenta los resultados obtenidos por el método *wdir* en todas las estaciones como hemos descrito en las tablas anteriores y realizando la media por contaminante, obtenemos la siguiente clasificación para nuestro sistema:

Contaminante	*BIAS*	*r*	*RMSE*	*Clasificación*
O_3	1.54 (MB)	0.89 (MB)	12.48 (B)	Entre Bueno y Muy Bueno
NO_2	-4.08 (MB)	0.82 (MB)	18.47 (A)	Entre Aceptable y Muy Bueno
SO_2	0.31 (MB)	0.48 (MB)	1.59 (MB)	Muy Bueno

Tabla 6.10: Clasificación del sistema de predicción según los valores de evaluación del sistema CALIOPE.

En general, según los resultados obtenidos en la predicción de los tres contaminantes teniendo en cuenta el tráfico según la dirección del viento y la clasificación calculada mediante la evaluación del sistema CALIOPE (tabla 6.10), la calidad de nuestro sistema es bastante buena: Muy buena en el caso del SO_2; entre Buena y Muy buena para el O_3; y entre Aceptable y Muy buena para el NO_2.

6.7 Ejemplo

En esta sección hemos obtenido un modelo caracterizado para cada contaminante en cada estación. En total, hemos creado 24 modelos (para 4 contaminantes en 6 estaciones).

Para ver el funcionamiento de estos modelos realizaremos una prueba de predicción de 24h en la estación de Molí del Sol. En concreto se usarán los datos para el día 25/08/2015 desde las 00h hasta las 23h[4].

En la tabla 6.11 se puede ver el dato real observado en la estación (*Obs.*), la predicción realizada por el modelo (*Pred.*) y el *Mean Error* (*ME*) calculado de la siguiente manera:

$$ME = Pred - Obs$$

Hora	NO			NO_2			O_3			SO_2		
	Obs.	*Pred.*	*ME(*)*	*Obs.*	*Pred.*	*ME(*)*	*Obs.*	*Pred.*	*ME(*)*	*Obs.*	*Pred.*	*ME(*)*
0	2	2.07	0.07	14	19.61	5.61	60	55.3	-4.7	3	3.14	0.14
1	2	2.12	0.12	32	19.4	-12.6	33	55.13	22.13	3	3.13	0.13
2	2	2.34	0.34	21	21.48	0.48	34	51.35	17.35	3	3.21	0.21
3	2	2.24	0.24	13	14.31	1.31	50	58.6	8.6	3	3.27	0.27
4	2	2.67	0.67	17	20.2	3.2	54	41.78	-12.22	3	3.36	0.36
5	2	3.51	1.51	16	21.31	5.31	55	40.19	-14.81	3	3.48	0.48
6	3	5.59	2.59	31	26.06	-4.94	39	44.93	5.93	3	3.69	0.69
7	7	9.12	2.12	40	34.14	-5.86	34	45.49	11.49	3	3.73	0.73
8	4	12.56	8.56	22	35.18	13.18	55	46.5	-8.5	3	3.51	0.51
9	4	5.08	1.08	18	23.59	5.59	71	72.78	1.78	3	3.31	0.31
10	4	4.79	0.79	16	22.61	6.61	81	76.96	-4.04	3	3.15	0.15
11	4	3.99	-0.01	14	17.49	3.49	83	82.36	-0.64	3	3.15	0.15
12	4	3.89	-0.11	14	16.19	2.19	89	86.99	-2.01	3	3.08	0.08
13	4	3.91	-0.09	14	14.68	0.68	87	91.7	4.7	3	3.12	0.12
14	3	3.66	0.66	12	14.53	2.53	90	88.38	-1.62	3	3.05	0.05
15	3	3.56	0.56	11	15.01	4.01	89	88.7	-0.3	3	3.06	0.06
16	3	3.27	0.27	12	13.99	1.99	91	88.97	-2.03	3	2.98	-0.02
17	4	2.72	-1.28	16	15.61	-0.39	96	86.93	-9.07	3	3.01	0.01
18	4	3.09	-0.91	24	20.38	-3.62	89	81.47	-7.53	3	3.01	0.01
19	2	2.82	0.82	16	18.37	2.37	83	80.59	-2.41	3	3.05	0.05
20	3	2.43	-0.57	18	23.61	5.61	77	82.02	5.02	3	3.05	0.05
21	3	2.28	-0.72	19	28.01	9.01	60	55.3	-4.7	3	3.03	0.03
22	2	2.03	0.03	12	15.4	3.4	33	55.13	22.13	3	3.02	0.02
23	2	1.96	-0.04	11	12.84	1.84	34	51.35	17.35	3	3	0

Tabla 6.11: Ejemplo de predicción 24h en la estación Molí del Sol. (*) Valor óptimo $ME = 0$.

[4]La selección de la fecha y la estación ha sido determinada en base al número de datos completos en el conjunto de *test*.

CAPÍTULO 7
Extrapolación de las predicciones

En la sección anterior hemos analizado como obtener predicciones de contaminación a partir de una serie de parámetros de intensidad del tráfico y climatológicos. Nuestro objetivo ahora es extrapolar estas predicciones a toda la ciudad, de modo que sea posible estimar los datos de contaminación en lugares donde físicamente no haya estación de medición.

7.1 Introducción

Como hemos visto en el capítulo 4, los métodos de interpolación espacial pueden estimar los valores en lugares donde no existen datos, utilizando valores conocidos en otros puntos. En esta sección se presentarán y probarán diferentes métodos de interpolación espacial con el fin de determinar aquel que obtenga los mejores resultados interpolando los niveles de contaminación.

7.1.1. Notación

A lo largo de este capítulo se describirán diferentes técnicas de interpolación de datos. Con el fin de que estas sean explicadas de manera comprensible y sus fórmulas puedan ser comparadas, se describirán todas mediante la siguiente notación:

- Supondremos una región o área geográfica S.
- Dispondremos de N valores conocidos, situados en los puntos $x_i \in S$, donde $i = 1, 2 \dots N$.
- El valor conocido en cada punto $x_i \in S$ se denotará como u_i, donde $i = 1, 2 \dots N$.
- Llamaremos x a un punto arbitrario dentro del espacio S del que queremos saber el valor, desconocido hasta el momento y que denotaremos como $u(x)$.
- La distancia geográfica entre el punto $x \in S$ y un punto $x_i \in S$ la definiremos como $d(x, x_i)$, donde $i = 1, 2 \dots N$.
- Algunos de los métodos descritos a continuación aplicarán un peso o ponderación a los valores, dependiendo de diferentes factores. El peso que se aplicará al valor u_i del punto $x_i \in S$ se denotará como w_i, donde $i = 1, 2 \dots N$.

7.2 Métodos de interpolación

Se han estudiado las siguientes técnicas de interpolación:

- **Media (*Mean*):** Un método de referencia donde siempre se predice la media de los N puntos conocidos.
$$u(x) = \frac{\sum_{i=1}^{N} u_i}{N}$$
- ***Inverse Distance Weighting* (IDW):** Los valores de los puntos no conocidos son calculados usando una media ponderada de los puntos conocidos. Los pesos son estimados usando las distancias al punto a predecir. En este trabajo se ha utilizado el conocido método de Shepard [83], con $p = 1$.
$$u(x) = \frac{\sum_{i=1}^{N} u_i w_i}{\sum_{i=1}^{N} w_i}$$
Donde $w_i = \frac{1}{d(x, x_i)^p}$.
- ***Local Inverse Distance Weighting* (LIDW):** Una versión modificada del método IDW. En esta versión asignamos una influencia mayor a los valores más cercanos al punto a interpolar comparado con IDW.
$$u(x) = \sum_{i=1}^{N} u_i w_i$$
Donde $w_i = \frac{D - d(x, x_i)}{D * (N-1)}$, siendo $D = \sum_{i=1}^{N} d(x, x_i)$.
- ***Wind Sensitive* LIDW:** Una modificación de LIDW que tiene en cuenta la dirección del viento, de manera que incrementa el peso de aquellos puntos a barlovento. Se define un sector circular $A = \pm 30^{o}$ desde el punto a predecir y los pesos de los puntos conocidos dentro del sector se incrementan por un factor $\alpha = 1{,}5$. Siendo ϱ el número de puntos dentro del sector y $\varphi = N - \varrho$, el número de puntos fuera, donde N es el número total de puntos conocidos, el valor del nuevo punto desconocido x se calcula del siguiente modo:
$$u(x) = \sum_{i=1}^{\varrho} u_i w_{i(\varrho)} w_{\varrho} + \sum_{i=1}^{\varphi} u_i w_{i(\varphi)} w_{\varphi}$$
El peso del punto $x_i \in \varrho$ (dentro del sector A) se calcula como $w_{i(\varrho)} = \frac{D_{\varrho} - d(x, x_i)}{D_{\varrho} * (\varrho - 1)}$, siendo $D_{\varrho} = \sum_{i=1}^{\varrho} d(x, x_i)$.
El peso del punto $X_i \in \varphi$ (fuera del sector A) se calcula como $w_{i(\varphi)} = \frac{D_{\varphi} - d(x, x_i)}{D_{\varphi} * (\varphi - 1)}$, siendo $D_{\varphi} = \sum_{i=1}^{\varphi} d(x, x_i)$.
En el caso de $\varrho = 1 \cap \varphi = 1$, el valor de w_i para el único punto $x_i \in \varrho \cap \varphi$ es 1.
En general, se cumple que $\sum_{i=1}^{\varrho} w_{i(\varrho)} = 1$ y $\sum_{i=1}^{\varphi} w_{i(\varphi)} = 1$.
El peso total para los puntos dentro del sector w_{ϱ} y el peso total para los puntos fuera del sector w_{φ} se calculan como sigue:
$$w_{\varrho} = (\alpha / N) * \varrho$$

$$w_\varphi = (\beta/N) * \varphi$$

Donde $\beta = \dfrac{N - (\alpha * \varphi)}{N - \varphi}$, cumpliendo que $w_\varrho + w_\varphi = 1$.

- **Kriging:** Kriging es un método óptimo de interpolación basado en regresión a partir de las observaciones de N valores de puntos en los alrededores, ponderados de acuerdo a valores de covarianza espacial. Existen tres variantes de Kriging (Simple, Ordinaria y Universal), pero todas se basan en una función básica de regresión lineal de $Z^*(x)$ definida como:

$$Z^*(x) - m(x) = \sum_{i=1}^{N} w_i[Z(x_i) - m(x_i)]$$

Donde $m(x)$ y $m(x_i)$ son lo valores esperados (medias) de $Z(x)$ y $Z(u_i)$. $Z(x)$ se calcula como un campo aleatorio con una tendencia (*trend*), $m(x)$ y un componente residual, $R(x) = Z(x) - m(x)$. Kriging estima el valor residual en x como la suma ponderada de los valores residuales de los datos de los puntos en los alrededores. Los pesos, w_i, se derivan de la función de covarianza o semivariograma [23][16]. La diferencia entre las tres variantes de Kriging radica en la asunción de que la media o tendencia es constante para todos los puntos, para los cercanos o solo para cada vector de coordenadas. Para problemas de estimación de concentraciones de contaminantes, tanto Kriging Ordinario como Kriging Universal obtienen resultados similares [49], sin embargo, Kriging Ordinario ha sido ampliamente implementado en paquetes de cálculos geoespaciales para R. Por este motivo, en este trabajo se ha decidido utilizar Kriging Ordinario. En concreto, se ha usado la implementación en R de la librería geoR [77], aunque también se ha probado la librería gstat [71] obteniendo los mismos resultados.

7.3 Experimentos

Dado que no tenemos datos de los niveles de contaminación de las zonas donde no existe estación de medición, no podemos comparar los resultados de la interpolación con datos reales. De la misma manera, solo disponemos de datos de seis estaciones, lo que limita el espacio geográfico de puntos conocidos. Para paliar estas limitaciones, primero se probará cada uno de los métodos de interpolación con los datos reales de las estaciones, suponiendo que una o más de ellas son puntos no conocidos e intentando obtener el valor de estas últimas a partir de los datos reales de las demás estaciones. Con los resultados de estas pruebas podremos saber qué método de los probados es el más adecuado para la interpolación de los contaminantes para, posteriormente, aplicarlo al resto de la ciudad (*grid*) a partir de los datos obtenidos en las seis estaciones.

Con la finalidad de evaluar los métodos de interpolación con los datos de las seis estaciones disponibles, se han establecido tres escenarios diferentes, donde se intenta predecir el valor de una o varias estaciones a partir de los valores de las demás mediante validación cruzada (Figura 7.1).

1. 3 estaciones como puntos conocidos, frente a 3 estaciones como puntos desconocidos (20 posibles combinaciones).
2. 4 estaciones como puntos conocidos, frente a 2 estaciones como puntos desconocidos (15 posibles combinaciones).
3. 5 estaciones como puntos conocidos, frente a 1 estación como punto desconocido (5 posibles combinaciones).

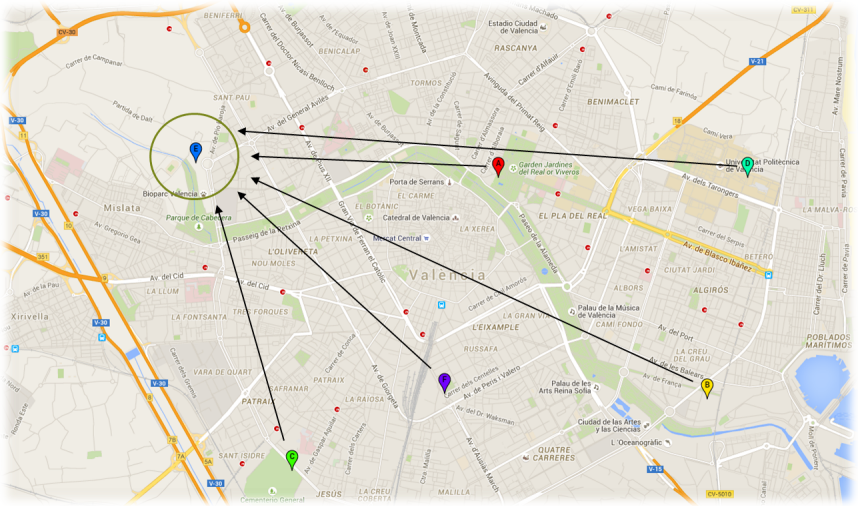

Figura 7.1: Ejemplo de validación cruzada para el escenario 3, con cinco estaciones con valores conocidos frente a una estación con valores desconocidos.

7.4 Resultados

	3 conocidos/3 desconocidos					4 conocidos/2 desconocidos					5 conocidos/1 desconocido				
	M	*I*	*L*	*W*	*K*	*M*	*I*	*L*	*W*	*K*	*M*	*I*	*L*	*W*	*K*
NO	22.2	17.7	19.6	**16.9**	19.9	18.1	17.9	17.8	**15.0**	19.9	17.1	14.8	16.7	**14.7**	23.0
NO_2	21.5	21.5	22.1	**20.0**	**20.0**	19.0	20.8	22.2	**17.3**	20.6	17.9	16.3	19.2	**16.1**	24.2
O_3	15.4	17.5	14.5	17.7	**13.2**	**12.3**	14.4	20.4	15.1	12.7	**12.0**	13.9	13.7	13.9	18.7
SO_2	3.1	3.0	3.2	2.9	**2.9**	**3.2**	3.0	3.0	**2.5**	3.4	3.0	**2.4**	2.9	2.5	3.4

Tabla 7.1: Comparación de los cinco métodos de interpolación espacial, para los cuatro contaminantes, en tres escenarios diferentes. RMSE de la predicción en los puntos desconocidos respecto a su valor real. La mejor predicción está resaltada en negrita. *M=Media; I=IDW; L=LIDW; W=Wind Sensitive LIDW; K=Kriging*

7.5 Discusión

La tabla 7.1 incluye la media de RMSE de los puntos no conocidos con respecto a su valor real, para los tres escenarios estudiados en los diferentes métodos de interpolación, para los cuatro contaminantes. En estos experimentos se usa el conjunto de datos completo de cada estación, es decir las medidas horarias de 2013, 2014 y 2015.

Si comparamos los resultados de los métodos de interpolación en esta tabla, Kriging y *Wind Sensitive* LIDW obtienen los mejores resultados. En general, Kriging interpola mejor cuando hay pocos valores conocidos y *Wind Sensitive* LIDW mejora cuando se añade más información a la interpolación. La excepción a este comportamiento es O_3, para el que los métodos de interpolación, en general, no consiguen mejorar el método básico. Esto puede ser debido a las características de este contaminante como hemos visto en el capítulo 5, que lo hacen mucho más independiente de la dirección del viento o la posición de la estación y mucho más dependiente del valor de otros contaminantes.

Dados estos resultados,*Wind Sensitive* LIDW y Kriging serán los métodos utilizados en el resto de esta memoria, a pesar de que la aplicación web que se mostrará en el capítulo 8 usará únicamente *Wind Sensitive* LIDW.

7.6 Ejemplo

Como hemos comentado anteriormente, no es posible comprobar los resultados de una extrapolación, puesto que no hay datos reales en los puntos a extrapolar. Sin embargo, dado que Kriging y *Wind Sensitive* LIDW son los métodos que cometen el menor error en la interpolación de los datos, vamos a utilizarlos para crear visualizaciones que hagan más comprensibles los resultados.

En este ejemplo extrapolaremos al resto de la ciudad los datos de contaminación de una hora aleatoria de las seis estaciones. En la tabla 7.2 podemos ver los niveles de contaminación que utilizaremos.

Estación	NO	NO_2	O_3	SO_2
Moli	42.82	51.04	32.25	3.60
Pista	34.90	50.18	35.49	4.72
Francia	25.65	43.89	42.98	3.43
Viveros	44.14	41.85	32.21	3.37
Bulevar	44.92	50.14	34.56	3.79
UPV	13.93	43.98	50.39	3.41

Tabla 7.2: Datos de contaminación utilizados para el ejemplo de extrapolación.

En concreto y para ver el resultado con varios contaminantes, utilizaremos los datos de NO_2 y O_3 con los dos métodos de interpolación.

Para realizar las visualizaciones que se mostrarán en las siguientes secciones, crearemos un *grid* de coordenadas, es decir, una cuadrícula sobre Valencia, de modo que cubramos toda la extensión de la ciudad y sea posible calcular el nivel de los contaminantes en cualquier zona. En este caso, nuestro *grid* contendrá al rededor de 22000 celdas, aumentando en 0.01 las coordenadas de cada una de ellas. Esto lo podemos hacer en R con la función expand.grid[1].

7.6.1. Kriging

En primer lugar realizaremos la extrapolación con Kriging, utilizando para ello la librería geoR y en particular, la función ksline con los parámetros explicados en [29].

Con esta función obtenemos un objeto que ya incluye la estimación del contaminante en cada celda del grid. Para generar una visualización con estos resultados utilizaremos la función image[2] e incluiremos como puntos sobre la imagen las coordenadas de las seis estaciones. En la figura 7.2 podemos ver el resultado obtenido para NO_2 (izquierda) y Ozono (derecha). Los colores indican la intensidad en el nivel de los contaminantes: cuánto más saturada está la imagen (más clara) el nivel del contaminante es menor.

Como se puede observar, el nivel del NO_2 es más bajo hacía el este, mientras que aumenta hacía el oeste. En el caso del Ozono ocurre lo contrario. Si comparamos estos resultados con los valores de la tabla 7.2, podemos ver que las estaciones situadas al este (UPV y Francia) corresponden a los valores más bajos de NO_2 (además de la estación de Viveros). El caso contrario lo vemos en el Ozono, para el que los valores de estas estaciones resultan ser los más altos.

[1] expand.grid: https://stat.ethz.ch/R-manual/R-devel/library/base/html/expand.grid.html
[2] image: https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/image.html

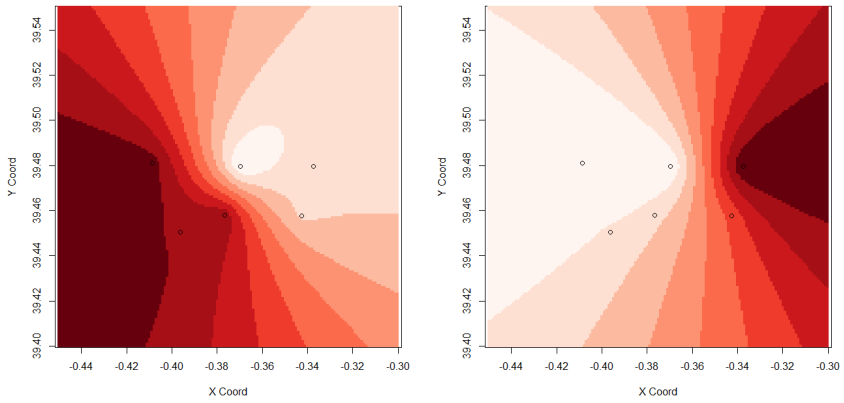

Figura 7.2: Ejemplo de extrapolación del NO_2 (izquierda) y el O_3 (derecha) generado mediante Kriging.

7.6.2. *Wind Sensitive* LIDW

En este caso vamos a realizar la extrapolación de los dos contaminantes con el método *Wind Sensitive* LIDW. Al ser un nuevo método, no dispone de una librería con la que generar los resultados, de modo que se usa la implementación que hemos utilizado anteriormente para realizar los experimentos.

Tal y como se ha visto en este capítulo, este método utiliza la dirección del viento para calcular los nuevos valores dependiendo de la posición de las estaciones. En este caso, el viento sopla del nordeste.

Para este segundo ejemplo de visualización vamos a generar un mapa de densidad sobre un mapa de Valencia, utilizando para ello la librería ggmap [50] de R. El resultado puede verse en la figura 7.3.

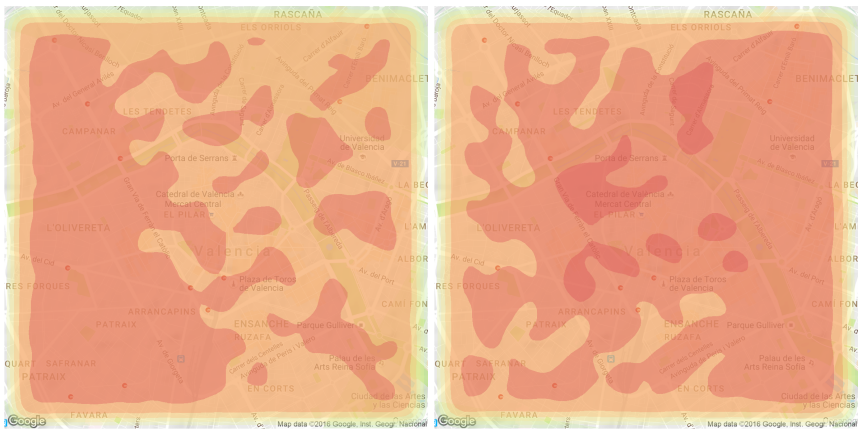

Figura 7.3: Ejemplo de extrapolación del NO_2 (izquierda) y el O_3 (derecha) generado mediante *Wind Sensitive* LIDW.

En este caso, donde la intensidad del color corresponde al nivel de concentración del contaminante, podemos volver a observar que los niveles más altos de NO_2 (izquierda) se encuentran hacía el oeste, mientras que los de O_3 (derecha) se distribuyen por el este de la ciudad. A diferencia de las imágenes creadas mediante Kriging, donde se observa un claro rango de niveles, en los mapas generados mediante nuestro método podemos ver como se dispersa o concentra la contaminación en función del viento.

CAPÍTULO 8
airVLC: Aplicación para visualizar los resultados

Para poder mostrar en funcionamiento los resultados de este proyecto se ha creado una aplicación web, airVLC, en la que se pueden consultar los últimos datos obtenidos, así como la predicción y extrapolación de las mismas mostradas visualmente en un mapa interactivo.

En esta sección describiremos el diseño e implementación de esta aplicación web.

8.1 Arquitectura del sistema

Nuestra aplicación está basada en el modelo Cliente/Servidor. Los clientes realizan peticiones al Servidor Web que le ofrece la respuesta, de este modo la capacidad de proceso está repartida entre los clientes y los servidores. Mediante esta arquitectura los accesos, recursos y la integridad de los datos son controlados por el servidor, de modo que un cliente no puede acceder a aquellos datos a los que no esté autorizado.

La arquitectura de nuestra web está estructurada de la siguiente manera:

- Nivel de presentación: Esta capa de nuestra arquitectura se encarga de la representación de la información para el usuario final, interactuando con él y comunicándose únicamente con el nivel de aplicación.
- Nivel de aplicación: Es donde se ubica el código de los programas, que se ejecutan, recibiendo las peticiones del usuario y enviándole las respuestas tras el proceso.
- Nivel de persistencia: En esta capa se encuentran los datos guardados y procesados por el nivel de aplicación.

8.1.1. Nivel de presentación

El nivel de Presentación está formado por todos los documentos que envía el servidor al cliente y que se muestran al usuario final de la aplicación, es decir, la interfaz gráfica. Esta web es un portal simple, sin intranet, con un menú superior desde el que se accede directamente a las diferentes secciones. A continuación se detalla cada una de ellas.

Figura 8.1: airVLC: Página principal.

Página principal

La página principal describe brevemente la finalidad del proyecto y los autores, y ofrece accesos directos, tanto con el menú superior (común en toda la web), como en los enlaces del cuerpo de la página (Figura 8.1).

Mapa

La página "Mapa"muestra un mapa interactivo creado con Leaflet con cinco capas disponibles, pudiendo superponerlas para ver todos los datos (Figura 8.2).

Por un lado, podemos ver la última predicción calculada mediante Random Forest para los cuatro contaminantes en las seis estaciones.

Por otro lado, tenemos cuatro capas de extrapolaciones, una por contaminante. Cada una de estas capas muestra 315 rectángulos (o celdas del *grid*), donde se puede ver la

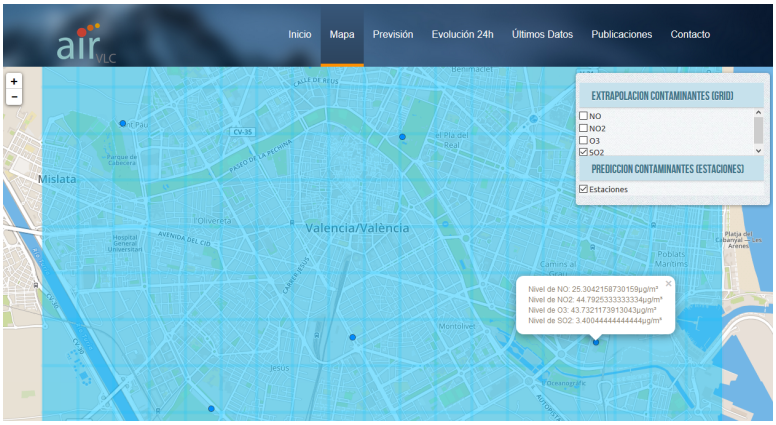

Figura 8.2: airVLC: Mapa.

extrapolación calculada mediante *Wind Sensitive LIDW* en cada una de estas zonas a partir de las predicciones anteriores.

En todas las capas, los colores (tanto de los puntos de las estaciones, como de las celdas del gris) cambiarán de color (de azul a rojo) según el nivel de los contaminantes, según se ha visto en el capítulo 2.

Evolución 24h

Figura 8.3: airVLC: Evolución 24h.

La sección de predicción muestra, para cada contaminante, una descripción del mismo y un gif (imagen animada) de la extrapolación mediante *Wind Sensitive LIDW* de las últimas 24h (Figura 8.3).

Últimos datos

En "Últimos datos"se pueden ver los últimos datos guardados en el sistema: la media de las últimas predicciones, los datos de la observación meteorológica y la media de tráfico en la ciudad (Figura 8.4).

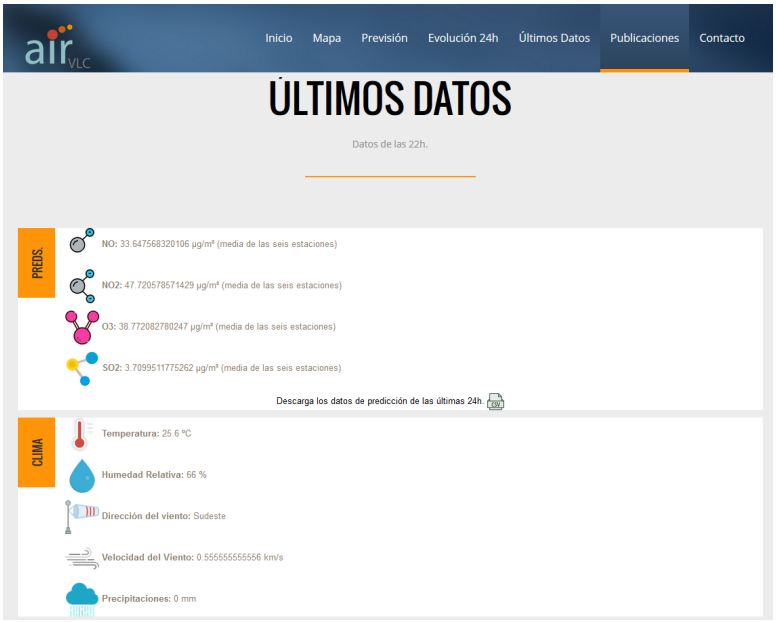

Figura 8.4: airVLC: Últimos datos.

En esta sección se pueden descargar las predicciones de las últimas 24 horas en formato CSV.

Publicaciones

Figura 8.5: airVLC: Publicaciones.

En "Publicaciones"se muestra el resumen del proyecto y se muestran las publicaciones relacionadas con este trabajo[1] (Figura 8.5).

8.1.2. Nivel de aplicación

Con esta capa definimos el comportamiento de los objetos que interactuarán en la web. Estos objetos contienen diversos tipos de funciones que se diferencian según su propósito: mostrar información, acceder a los datos o procesar datos.

[1] Las publicaciones relacionadas con este trabajo pueden verse en el apéndice A.

En la figura 8.6, se muestran las diferentes acciones que se desarrollan en el sistema y que se describirán con más detalle en la sección 8.4.

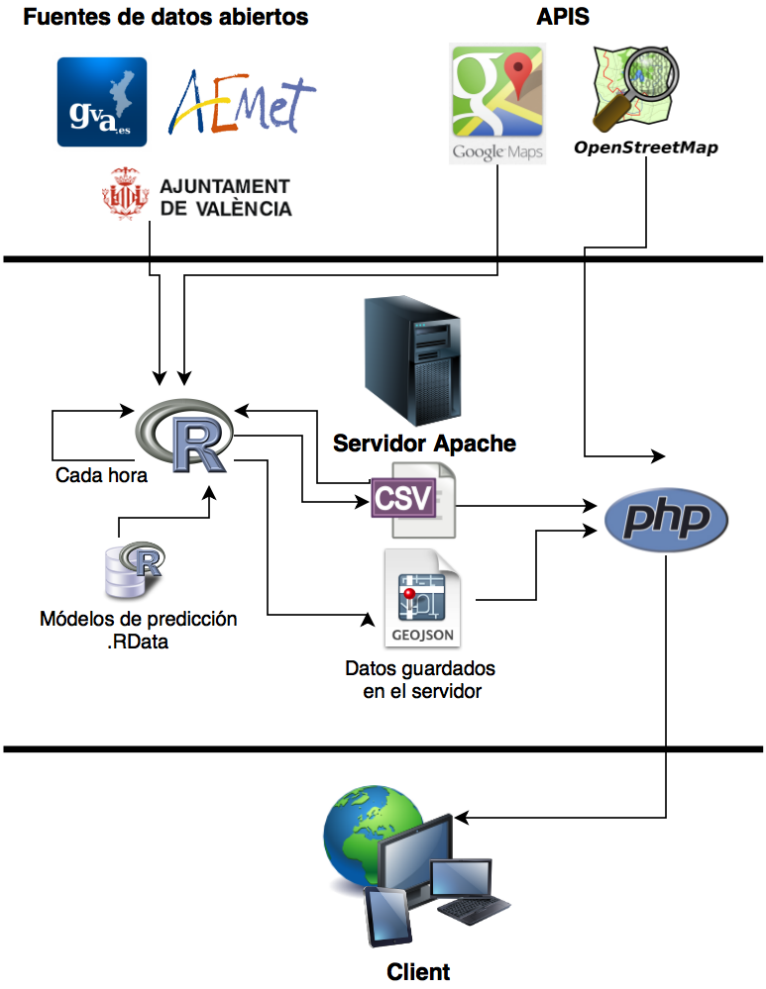

Figura 8.6: airVLC: Arquitectura del sistema.

8.1.3. Nivel de persistencia

Es en esta capa donde los datos de nuestra aplicación están almacenados. En nuestro caso no se ha implementado base de datos, si no que los datos son almacenados en ficheros CSV o GeoJSON según su propósito.

Los datos se almacenan por horas, en carpetas separadas para cada una de ellas. En cada una de estas carpetas se guardan en primer lugar y en formato CSV, los siguientes datos:

- Observación meteorológica.
- Intensidad del tráfico en todas las espiras disponibles.
- Predicción para los cuatro contaminantes en las seis estaciones.
- Datos de la extrapolación de los cuatro contaminantes para las 315 celdas del *grid*.

A continuación y en formato geoJSON se vuelven a guardar los resultados para la extrapolación, en este caso con las coordenadas de cada polígono del *grid*.

8.2 Tecnologías utilizadas

Para realizar esta web se han utilizado diferentes tecnologías y lenguajes. Principalmente, se ha utilizado PHP para crear el portal; HTML y CSS para el aspecto visual; jQuery (JavaScript) para controlar el comportamiento de la web en el lado del cliente; R en el lado del servidor para generar las predicciones y por último, CSV y GeoJSON para almacenar los resultados.

A continuación se describe cada una de estas tecnologías.

8.2.1. HTML

El HTML, *Hyper Text Markup Language* (Lenguaje de marcación de Hipertexto) es un estándar utilizado en la www (World Wide Web) para crear páginas web y aplicaciones. Fue creado en 1986 por el físico nuclear Tim Berners-Lee a partir del concepto de Hipertexto (Conocido también como link o ancla) y el SGML (Lenguaje Estándar de Marcación General) que sirve para colocar etiquetas o marcas en un texto para indicar como debe verse. HTML es un sistema de etiquetas que no requiere compilador[2].

8.2.2. CSS

Hojas de Estilo en Cascada (*Cascading Style Sheets*), es un mecanismo que describe cómo se va a mostrar un documento en la pantalla, o cómo se va a imprimir, o incluso cómo va a ser pronunciada la información presente en ese documento a través de un dispositivo de lectura. CSS se utiliza para dar estilo a documentos HTML y XML, separando el contenido de la presentación en base a reglas[3].

8.2.3. PHP

PHP (*Hypertext Pre-processor*) es un lenguaje de programación interpretado del lado del servidor, diseñado originalmente por Rasmus Lerdorf en 1995 para la creación de páginas web dinámicas. Puede ser desplegado en la mayoría de los servidores web y en casi todos los sistemas operativos y plataformas[4].

8.2.4. JavaScript

JavaScript es un lenguaje de programación interpretado, utilizado normalmente en el lado del cliente, implementado como parte de un navegador web permitiendo mejoras en la interfaz de usuario y páginas web dinámicas. Para interactuar con una página web se provee al lenguaje JavaScript de una implementación del *Document Object Model* (DOM)[5].

jQuery

jQuery es una biblioteca o *framework* de JavaScript, creada inicialmente por John Resig, que permite simplificar la manera de interactuar con los documentos HTML, ma-

[2]HTML. W3C: https://www.w3.org/html/
[3]Hojas de estilo (CSS). Wc3: http://www.w3c.es/Divulgacion/GuiasBreves/HojasEstilo
[4]PHP: http://php.net/
[5]JavaSript: http://www.w3schools.com/js/

nipular el árbol DOM, manejar eventos, desarrollar animaciones y agregar interacción mediante AJAX a páginas web[6].

Leaflet

Leaflet es una librería de JavaScript de código abierto usada para construir mapas interactivos. Está soportada por la mayoría de los navegadores web y móviles, ya que está creada para ser usada con HTML5 y CSS3. Permite utilizar mapas de OpenLayers[7], OpenStreetMaps[8] y Google Maps API[9], entre otros[10].

8.2.5. CSV

Los archivos CSV (*comma-separated values*) son documentos en formato abierto en forma de tabla, donde las columnas se separan normalmente por comas y las filas por saltos de línea.

8.2.6. JSON

JSON, acrónimo de *JavaScript Object Notation*, es un formato estandar ligero para el intercambio de datos, consistente en pares de objetos atributo-valor. JSON es un subconjunto de la notación literal de objetos de JavaScript[11].

GeoJSON

GeoJSON es un formato estándar abierto, basado en JSON, para representar elementos geográficos. La gramática está basada en el estándar WKT (*Well Known Text*) del *Open Geospatial Consortium*[12], permitiendo diferentes tipos de geometrías (puntos, líneas, polígonos y colecciones).

8.3 Herramientas utilizadas

Para la implementación de la web se han utilizado diferentes aplicaciones o herramientas, que son las siguientes:

- Para la creación de los Scripts de PHP, que incluyen JavaScript y HTML y para las hojas de estilos (CSS) se ha utilizado Notepad++.
- Para las pruebas en un servidor Apache local se ha utilizado el paquete XAMPP.
- Para las pruebas y comprobaciones de la web se han utilizado distintos navegadores (Firefox, Internet Explorer y Chrome) sobre los sistemas operativos Windows y Android con diferentes resoluciones, así como un complemento para los navegadores llamado Firebug para la comprobación de errores de JavaScript y el desarrollo del CSS de la web.

[6] jQuery: https://jquery.com/
[7] OpenLayers: http://openlayers.org/
[8] OpenStreetMap: http://www.openstreetmap.org
[9] Google Maps API: https://developers.google.com/maps/?hl=es
[10] Leaflet: http://leafletjs.com/
[11] JSON. Wikipedia: http://es.wikipedia.org/wiki/JSON
[12] Open Geospatial Consortium: http://www.opengeospatial.org/

8.4 Implementación detallada

A continuación se mostrará una descripción más detallada de las partes más significativas de la implementación del proyecto (la generación de los datos y la visualización de los mismos. No obstante, todo el código está publicado en Github[13].

8.4.1. airVLC.R

Este *script* en R está programado para ejecutarse cada hora y es el encargado de ejecutar los modelos y métodos estudiados en las secciones anteriores con el fin de obtener tanto la predicción de los contaminantes, como la extrapolación de los mismos a toda la ciudad.

En principio, este script se creó para trabajar con datos guardados en una base de datos MySQL, sin embargo y debido a la gran cantidad de información que se tenía que manejar al generar los mapas de densidad, se vio que R trabajaba más rápido leyendo y escribiendo los datos en archivos CSV o JSON que realizando consultas a la base de datos. Por esta razón y pese a la perdida de disponibilidad e integridad de los datos en el apartado web, se decidió continuar esta aplicación mediante el uso de este tipo de archivos.

Datos y parámetros

En la página web de datos online de la Generalitat Valenciana, es posible consultar los últimos datos publicados sobre mediciones de contaminación en cada estación de la Comunidad Valenciana. Sin embargo, además de la demora de tres horas con la que se publican, estos datos no se encuentran en un formato abierto o reutilizable, si no que se comparten en forma de marcos y tablas HTML generadas con JAVA. Esto supone un problema a la hora de automatizar la utilización de estos datos en tiempo real en la aplicación. Debido a esto y teniendo en cuenta un aumento del error en las predicciones, para la realización de este script se ha prescindido de los datos observados tres horas antes con los que si se estudiaron los modelos en el capitulo 6.

Por lo tanto, los datos de entrada para este *script* son los siguientes:

- Número, nombre y coordenadas de las estaciones de medición de contaminación.
- Número y nombre de los contaminantes.
- Hora del sistema.
- Datos de tráfico en tiempo real, descargados de la web de datos abiertos del Ayuntamiento de Valencia en formato CSV.
- Datos de la observación meteorológica en tiempo real, descargados de la web de AEMET en formato CSV[14] [15].
- Fichero CSV creado previamente que contiene las espiras cercanas (1km) a cada una de las estaciones y sus coordenadas.

[13] https://github.com/liconoc/airVLC

[14] Datos de la observación meteorológica: http://www.aemet.es/es/eltiempo/observacion/ultimosdatos?k=val&l=8416Y&w=0&datos=det

[15] Nota: En estos datos la velocidad del viento viene en km/h, de modo que se realiza la conversión a m/s, tal y como se utilizan en los modelos al usar los datos de la GVA, según se ha visto en los capítulos 5 y 6

Una vez se han descargado todos los datos se guardan en archivos CSV, en la carpeta correspondiente a la hora del sistema, tanto los datos de tráfico como la observación meteorológica para esa hora.

Predicción

Para realizar la predicción se utilizan los datos descargados y el modelo previamente guardado en formato RData y entrenado con los datos de los tres años de históricos (uno por contaminante para cada estación). El algoritmo 1 muestra en pseudocódigo el funcionamiento de esta parte del *script*.

Algorithm 1 airVLC - Predicción

```
1: for all estaciones do
2:     puntosCercanos ← espiras < 1km
3:     if existe direccionViento then                    ▷ Si no está en calma
4:         TRANSFORMARDIRECCIONVIENTO(direccionViento)   ▷ Convertir al formato del rumbo
5:     end if
6:     CALCULARPUNTOSDENTRO(direccionViento, puntosCercanos)   ▷ Determinar las espiras a barlovento
7:     for all contaminantes do
8:         CARGARMODELO(contaminante, estacion)
9:         prediccion ← PREDICCION(datosContaminante)   ▷ Se aplica el modelo con los datos descargados
10:    end for
11: end for
12: GUARDARCSV(prediccion)
```

Extrapolación

La segunda parte de la aplicación calcula la extrapolación de las predicciones al resto de la ciudad mediante un *grid*. Esto se hace de dos formas diferentes:

1. La primera forma es creando 315 polígonos que cubren la ciudad de norte a sur y de este a oeste. La extrapolación en cada celda o polígono se calcula mediante las coordenadas centrales del mismo. Los resultados se guardan en un archivo GeoJSON que será utilizado por el mapa de la web.

2. 2. La segunda forma genera también un *grid*, pero esta vez de aproximadamente 22000 celdas, como se ha visto en el capítulo X. De esta manera cada zona a extrapolar es mucho más pequeña, de modo que la precisión es más alta. Los resultados de esta forma se utilizan para generar las imágenes y los gifs[16] para la web.

El pseudocódigo de la extrapolación, específicamente de la primera forma, puede verse en el algoritmo 2.

[16]Los gifs se generan mediante la librería de R [93].

Algorithm 2 airVLC - Extrapolación

```
1: CrearGrid(coordenadasLimite, distanciasCeldas)
2: for all contaminantes do
3:     for all celda do
4:         coordenadas ← coordenadasCentralesGrid
5:         matrizDistancias ← calcularMatrizDistancias(coordenadas, estaciones)
6:         if existe direccionViento then                    ▷ Si no está en calma
7:             TransformarDireccionViento(direccionViento)   ▷ Convertir al formato del rumbo
8:         end if
9:         puntosDentro ← CalcularPuntosDentro(direccionViento, matrizDistancias)
   ▷ Determinar las estaciones a barlovento
10:        pesos ← CalcularPesos(matrizDistancias)
11:        extrapolacion ← CalcularExtrapolacion(pesos, predicciones)
12:    end for
13:    GuardarGeoJson(extrapolacion)
14: end for
```

8.4.2. mapa.php

Esta sección de la web muestra el mapa con los datos de la predicción y la extrapolación de los contaminantes. Para llevar a cabo esta tarea se utilizan principalmente dos tecnologías: PHP para cargar los datos y JavaScript para generar el mapa.

La librería Leaflet proporciona las herramientas necesarias para generar un mapa interactivo en cualquier sección de una página web. En nuestro caso, mostramos un mapa centrado en la ciudad de Valencia.

```
var map = L.map('map', { zoomControl:true, scrollWheelZoom:false
    }).setView([39.4532093,-0.3783341], 13);
```

Además de esto, esta herramienta nos permite personalizar los mapas de diversos modos, incluyendo el poder cambiar los iconos de los puntos en el mapa. De este modo, las estaciones están identificadas mediante círculos del color correspondiente al índice de calidad del aire, y los polígonos se dibujan como rectángulos semitransparentes.

Las diferentes capas que se pueden superponer se generan a partir de un *plugin* para Leaflet, *Styled Layer Control*[17], que nos permite agrupar los puntos o polígonos en diferentes secciones y seleccionar cuáles mostrar.

Para esta sección de la web el uso de PHP se limita a la carga de archivos de datos en CSV y el paso de parámetros a JavaScript.

Los archivos GeoJSON, sin embargo, se cargan mediante la función getJSON de jQuery.

```
$.getJSON( "ficheros/datos/<?=$hora?>/grid_<?=$contaminante?>.geojson",
    function( data ) {...}
```

[17] *Styled Layer Control*: https://github.com/davicustodio/Leaflet.StyledLayerControl

CAPÍTULO 9
Conclusiones

La mala calidad del aire es un factor que puede disminuir la esperanza de vida, dado que la contaminación aumenta el riesgo de sufrir enfermedades respiratorias. A su vez, la contaminación es capaz de degenerar los ecosistemas y el medio ambiente en general. Existen sensores capaces de medir los niveles de los contaminantes, sin embargo y debido al coste que supone, el número de estas estaciones de medición es limitado y además de esto, el tiempo que algunos de estos contaminantes necesitan para ser validados supone un retraso considerable en la publicación de los datos. La detección de los niveles de contaminación altos en tiempo real, puede reducir la exposición a los ambientes contaminados y los efectos adversos que produce en la salud de las personas y el deterioro de la vida en general, además de suponer una herramienta útil que debe ser considerada para poder controlar los niveles de contaminación emitidos.

En este trabajo hemos utilizado datos históricos de Valencia, siendo esta ciudad nuestro caso de estudio, con el fin de encontrar aquella información que fuera de utilidad para poder predecir los niveles de contaminación. de modo que esta información pudiera ser publicada con antelación al dato real observado. Para ello, se han seleccionado diferentes fuentes de datos abiertos y aquellos parámetros que, después de su estudio, se ha viso que tenían suficiente relación con la contaminación para poder ser utilizados en los experimentos. En concreto, se han seleccionado diferentes parámetros meteorológicos y los datos de la intensidad del tráfico en Valencia para predecir el nivel de cuatro contaminantes (NO, NO_2, O_3, SO_2).

Con los datos seleccionados y los conjuntos de datos creados, hemos estudiado métodos de aprendizaje automático que pueden predecir los niveles de los contaminantes en seis estaciones de medición de contaminación en la ciudad de Valencia. De acuerdo con nuestros experimentos, *Random Forest* es la técnica capaz de generar las mejores predicciones en la mayoría de los casos estudiados. Hemos estudiado cómo mejorar estos modelos incorporando la dirección del viento, debido a que es un factor clave en la dispersión de los contaminantes. Se ha propuesto una aproximación donde la dirección del viento es usada para seleccionar dinámicamente las fuentes de emisión de tráfico. Los resultados muestran que esta aproximación es capaz de mejorar los resultados de las predicciones, para aquellos contaminantes directamente relacionados con las emisiones procedentes de la combustión de carburantes.

Posteriormente, se han estudiado diferentes técnicas de interpolación, con el fin de encontrar aquella que pudiera extrapolar de manera más precisa los datos de contaminación de las seis estaciones a toda la ciudad. Para este propósito, hemos presentado un nuevo método de interpolación, *Wind Sensitive* LIDW, que tiene en cuenta la dirección del viento a la hora de ponderar los valores de los puntos a interpolar. Hemos comparado esta nueva técnica con otras más conocidas como Kriging o IDW. Los experimentos

muestran que nuestro nuevo método, obtiene una mejora positiva en los resultados, especialmente cuando hay un número mayor de puntos conocidos para usar en la interpolación.

Finalmente, se ha generado un portal web, airVLC, que muestra en un mapa interactivo tanto la predicción de los contaminantes en las estaciones, como la extrapolación de estas predicciones a toda la ciudad. Esto se realiza cada hora, utilizando datos de actualización constante, proporcionados por diferentes organismos de manera abierta. Los resultados mostrados en la web, aunque con un margen de error dependiente del contaminante, son capaces de publicarse con antelación a los datos reales observados, proporcionando una idea de los niveles de contaminación existentes en la ciudad en cualquier momento.

9.1 Trabajo futuro

Como trabajo futuro se plantean una serie de retos que resuelvan las incertidumbres que han surgido en los resultados y/o que mejoren los modelos y técnicas utilizadas y en consecuencia, los resultados.

Dado que cada estación se encuentra ubicada en una zona diferente y esto, posiblemente, es la causa de la variedad de resultados en la predicción, se propone la aplicación de nuevas características locales en los puntos a predecir, por ejemplo, la altitud o la topografía. De igual manera, se puede estudiar el comportamiento de cada contaminante en las diferentes estaciones, ya que las ubicaciones o tipos de zona (centro de la ciudad, extrarradio...) y sus características (muchos edificios, zonas verdes...) pueden afectar al origen/dispersión de la contaminación. Otra opción sería estudiar la relación de nuevos parámetros que no se han tenido en cuenta como, por ejemplo, la radiación solar, que afecta tanto a la creación del Ozono como a la eliminación de los NO_x.

Se puede estudiar incorporar a los modelos de predicción la velocidad del viento, no como parámetro, si no como una característica a la hora de seleccionar los puntos de tráfico, de igual forma que la dirección del viento, ya que dependiendo de la velocidad del viento, los puntos de emisión que afectan a una estación de medición de contaminación pueden ser pocos y más cercanos o más lejanos y en mayor cantidad. Para mejorar el modelo de predicción dinámicamente, se puede incorporar el Filtro de Kalman [51], cuyo objetivo es corregir los errores cometidos por el modelo realimentándolo con los datos reales cuando se encuentran disponibles.

Por otro lado, se propone aplicar estas técnicas en otras ciudades para estudiar si el comportamiento es similar en la predicción/interpolación de los datos. Con este propósito se sugiere el uso de técnicas de *transfer learning* [69], con el fin de reutilizar el conocimiento adquirido en este proyecto con los datos generados en otras ciudades.

Como mejoras para la web, se propone el uso de una fuente de datos diferente para los datos meteorológicos, ya que los datos de AEMET pueden fallar a veces y dejar de estar disponibles. En su lugar, se pueden usar fuentes de datos abiertas como OpenWeatherMap[1]. De igual manera, se propone el estudio de nuevos métodos de creación de mapas de calor, densidad o temáticos para mejorar las visualizaciones de las predicciones [61]. Por último y como se ha visto en el capítulo 8, para poder usar los datos de los contaminantes publicados online por la GVA con tres horas de retraso, se propone el uso de técnicas de *web scraping* [35] para extraer la información directamente de la web, con el fin de mejorar los modelos online de airVLC.

[1]OpenWeatherMap: http://openweathermap.org/

Bibliografía

[1] Air pollution is sending tiny magnetic particles into your brain | New Scientist. https://www.newscientist.com/article/2104654-air-pollution-is-sending-tiny-magnetic-particles-into-your-brain/.

[2] Conceptos de minería de datos. https://msdn.microsoft.com/es-es/library/ms174949.aspx#DefiningTheProblem.

[3] CRISP-DM, still the top methodology for analytics, data mining, or data science projects. http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[4] Métodos de interpolación a partir de puntos. universidad de murcia. http://www.um.es/geograf/sigmur/temariohtml/node43_mn.html.

[5] Multiple Linear Regression. http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm.

[6] OMS | Calidad del aire (exterior) y salud. http://www.who.int/mediacentre/factsheets/fs313/es/.

[7] Principales contaminantes presentes en la atmósfera - Generalitat Valenciana. http://www.habitatge.gva.es/web/calidad-ambiental/principales-contaminantes-presentes-en-la-atmosfera.

[8] Tipos de modelos atmosféricos. http://www.inecc.gob.mx/calaire-herramientas-analisis/582-calaire-tipos-modelos.

[9] Transporte terrestre. http://mediambient.gencat.cat/es/05_ambits_dactuacio/atmosfera/la_contaminacio_atmosferica/fonts_demissio_de_contaminants/transport_terrestre/.

[10] WHO | WHO Global Urban Ambient Air Pollution Database (update 2016). http://www.who.int/phe/health_topics/outdoorair/databases/cities/en/.

[11] Halit Apaydin, F Kemal Sonmez, and Y Ersoy Yildirim. Spatial interpolation techniques for climate data in the gap region in turkey. *Climate Research*, 28(1):31–40, 2004.

[12] M.A. Arain, R. Blair, N. Finkelstein, J.R. Brook, T. Sahsuvaroglu, B. Beckerman, L. Zhang, and M. Jerrett. The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. *Atmospheric Environment*, 41(16):3453 – 3464, 2007.

[13] Ferran Ballester Díez, José María Tenías, and Santiago Pérez-Hoyos. Efectos de la contaminación atmosférica sobre la salud: una introducción. *Revista española de salud pública*, 73(2):109–121, 1999.

[14] Paul E Benson. Caline3-a versatile dispersion model for predicting air pollutant levels near highways and arterial streets. interim report. Technical report, California State Dept. of Transportation, Sacramento (USA). Transportation Lab., 1979.

[15] Roger Bivand, Tim Keitt, and Barry Rowlingson. rgdal: Bindings for the geospatial data abstraction library. *R package version 0.8-16*, 2014.

[16] Geoff Bohling. Kriging. *Kansas Geological Survey, Tech. Rep*, 2005.

[17] Marco Borga and Andrea Vizzaccaro. On the interpolation of hydrologic variables: formal equivalence of multiquadratic surface fitting and kriging. *Journal of Hydrology*, 195(1):160–171, 1997.

[18] Richard W Boubel, Daniel Vallero, Donald L Fox, Bruce Turner, and Arthur C Stern. *Fundamentals of air pollution*. Elsevier, 2013.

[19] L Breiman, JH Friedman, RA Olshen, and CJ Stone. Classification and regression trees, wadsworth international group, belmont, california, usa, 1984; bp roe et al., boosted decision trees as an alternative to artificial neural networks for particle identificatio n. *Nucl. Instrum. Meth. A*, 543:57, 2005.

[20] Jason Brownlee. A Tour of Machine Learning Algorithms, November 2013.

[21] Eugene Brusilovskiy. A Brief Introduction to Spatial Interpolation. http://www.bisolutions.us/A-Brief-Introduction-to-Spatial-Interpolation.php.

[22] Vidal Domínguez M.J. y Moreno Jiménez A. Cañada Torrecilla, R. Interpolación espacial y visualización cartográfica para el análisis de la justicia ambiental: ensayo metodológico sobre la contaminación por partículas atmosféricas en madrid. *Tecnologías de la Información Geográfica: La Información Geográfica al servicio de los ciudadanos*.

[23] Roberto Carletti, Maria Picci, and Daniela Romano. Kriging and bilinear methods for estimating spatial pattern of atmospheric pollutants. *Environmental monitoring and assessment*, 63(2):341–359, 2000.

[24] David C. Carslaw and Karl Ropkins. openair — an r package for air quality data analysis. *Environmental Modelling and Software*, 27–28(0):52–61, 2012.

[25] DC Carslaw. The openair manual–open-source tools for analysing air pollution data. *Manual for version*, 1:1–1, 2012.

[26] Alexandre Champendal, Mikhail Kanevski, and Pierre-Emmanuel Huguenot. Air pollution mapping using nonlinear land use regression models. In *International Conference on Computational Science and Its Applications*, pages 682–690. Springer, 2014.

[27] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000.

[28] Colin Chen. An introduction to quantile regression and the quantreg procedure. In *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. SAS Institute Inc. Cary, NC, 2005.

[29] Nicolas Christou. Ordinary kriging using geor and gstat.

[30] Alan J Cimorelli, Steven G Perry, Akula Venkatram, Jeffrey C Weil, Robert J Paine, Robert B Wilson, Russell F Lee, Warren D Peters, and Roger W Brode. Aermod: A dispersion model for industrial source applications. part i: General model formulation and boundary layer characterization. *Journal of Applied Meteorology*, 44(5):682–693, 2005.

[31] Lidia Contreras and Cèsar Ferri. Wind-sensitive interpolation of urban air pollution forecasts. *Procedia Computer Science*, 80:313 – 323, 2016. International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA.

[32] Barcelona Supercomputing Center Departamento de Ciencias de la Tierra. EvaluaciÓn del sistema de pronÓstico de calidad del aire caliope en espaÑa 2013.

[33] Peter A Eckhoff and Thomas N Braverman. Addendum to the user's guide to cal3qhc version 2.0 (cal3qhcr user's guide. 1995.

[34] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[35] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques: a survey. *Knowledge-based systems*, 70:301–323, 2014.

[36] Peter Flach. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, 2012.

[37] F Follos Pliego. Lenguaje r aplicado al análisis de datos de calidad del aire. manual básico para el tratamiento de datos de calidad del aire mediante el lenguaje estadístico r y paquetes adicionales como openair, 2012.

[38] Cambio global España. 2020/2050. energía. *Economía y Sociedad*, 2011.

[39] Cristina Guerreiro, Frank de Leeuw, Valentin Foltescu, et al. Air quality in europe-2013 report. *European Environment Agency Rep*, 2013.

[40] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[41] Steven R Hanna, Lloyd L Schulman, Robert J Paine, Jonathan E Pleim, and Mitchell Baer. Development and evaluation of the offshore and coastal dispersion model. *Journal of the Air Pollution Control Association*, 35(10):1039–1047, 1985.

[42] José Hernández Orallo, Maria Jóse Ramírez Quintana, and César Ferri Ramírez. *Introducción a la Minería de Datos*. Pearson Prentice Hall, 2004.

[43] Jon M Heuss, Dennis F Kahlbaum, and George T Wolff. Weekday/weekend ozone differences: what can we learn from them? *Journal of the Air & Waste Management Association*, 53(7):772–788, 2003.

[44] Robert J Hijmans, Ed Williams, Chris Vennes, and Maintainer Robert J Hijmans. Package 'geosphere', 2015.

[45] Gerard Hoek, Rob Beelen, Kees de Hoogh, Danielle Vienneau, John Gulliver, Paul Fischer, and David Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33):7561 – 7578, 2008.

[46] Stijn Janssen, Gerwin Dumont, Frans Fierens, and Clemens Mensink. Spatial interpolation of air pollution measurements using {CORINE} land cover data. *Atmospheric Environment*, 42(20):4884 – 4903, 2008.

[47] Kevin Johnston, Jay M Ver Hoef, Konstantin Krivoruchko, and Neil Lucas. *Using ArcGIS geostatistical analyst*, volume 380. Esri Redlands, 2001.

[48] Lucía Juan Montesinos M. Mercedes Tomás Tomás José Diéguez Rodríguez Enrique Mantilla Iglesias Miguel Poquet Peiró José V. Miró Bayarri, Rafael Orts Bargues. *La calidad del aire en la Comunidad Valenciana 2002, 2003, 2004*. Generalitat Valenciana, Consellería de Territori i Habitatge, 2005.

[49] Andre G Journel and ME Rossi. When do we need a trend model in kriging? *Mathematical Geology*, 21(7):715–739, 1989.

[50] David Kahle and Hadley Wickham. ggmap: Spatial visualization with ggplot2. *R Journal*, 5(1), 2013.

[51] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[52] Mikhail Kanevski and Michel Maignan. *Analysis and modelling of spatial environmental data*, volume 6501. EPFL press, 2004.

[53] A Karppinen, J Kukkonen, T Elolahde, M Konttinen, and T Koskentalo. A modelling system for predicting urban air pollution:: comparison of model predictions with the data of an urban measurement network in helsinki. *Atmospheric Environment*, 34(22):3735–3743, 2000.

[54] Klea Katsouyanni, Joel Schwartz, C Spix, Giota Touloumi, D Zmirou, A Zanobetti, B Wojtyniak, JM Vonk, A Tobias, A Ponka, et al. Short term effects of air pollution on health: a european approach using epidemiologic time series data: the aphea protocol. *Journal of epidemiology and community health*, 50(Suppl 1):S12–S18, 1996.

[55] Mukesh Khare and SM Shiva Nagendra. *Artificial neural networks in vehicular pollution modelling*, volume 41. Springer, 2006.

[56] Roger Koenker. *quantreg: Quantile Regression*, 2015. R package version 5.11.

[57] Lukasz A Kurgan and Petr Musilek. A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(01):1–24, 2006.

[58] Jin Li and Andrew D Heap. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. *Ecological Informatics*, 6(3):228–241, 2011.

[59] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[60] P Amparo López Jiménez and Vincent Espert Alemany. *Dispersión de contaminantes en la atmósfera*. Universidad Politécnica de Valencia, Servicio de Publicaciones, 2000.

[61] Robin Lovelace and James Cheshire. Introduction to visualising spatial data in r. 2014.

[62] Ernesto Martínez Ataz and Yolanda Díaz de Mera Morales. *Contaminación atmosférica*. Ediciones de la Universidad de Castilla-La Mancha, 2004.

[63] Laina D Mercer, Adam A Szpiro, et al. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (nox) for the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Atmospheric Environment*, 45(26):4412–4420, 2011.

[64] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. R package version 1.6-4.

[65] Alimentación y Medio Ambiente Ministerio de Agricultura. Informe de la calidad del aire en españa 2014. http://www.magrama.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/informeevaluacioncalidadaireespana2014_final_tcm7-398522.pdf, 2015.

[66] Subdirección General de Calidad del Aire y Medio Ambiente Industrial Ministerio de Agricultura, Alimentación y Medio Ambiente. Análisis de la calidad del aire en españa: Evolución 2001 a 2012. http://www.isciii.es/ISCIII/es/contenidos/fd-el-instituto/fd-organizacion/fd-estructura-directiva/fd-subdireccion-general-servicios-aplicados-formacion-investigacion/fd-centros-unidades/fd-centro-nacional-sanidad-ambiental/fd-servicios-cientifico-tecnicos_sanidad-ambiental/Analisis_calidad_aire_Espana_2001_2012_tcm7_311112.pdf, 2013.

[67] Lubos Mitas and Helena Mitasova. Spatial interpolation. *Geographical information systems: principles, techniques, management and applications*, 1:481–492, 1999.

[68] Lidia Contreras Ochando, Cristina I. Font Julián, Francisco Contreras Ochando, and Cèsar Ferri Ramirez. Airvlc: An application for real-time forecasting urban air pollution. In *Proceedings of the 2nd International Workshop on Mining Urban Data*, pages 72–79, 2015.

[69] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[70] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[71] Edzer J Pebesma. Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, 30(7):683–691, 2004.

[72] Steven G Perry. Ctdmplus: A dispersion model for sources near complex topography. part i: Technical formulations. *Journal of Applied Meteorology*, 31(7):633–645, 1992.

[73] Las Provincias. *Enciclopedia Universal Las Provincias*. Domenech S.A., 1990.

[74] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[75] Redacción. El impactante hallazgo sobre la contaminación en cerebros de personas que vivieron y murieron en Ciudad de México. *BBC Mundo*, September 2016.

[76] Z Reitermanova. Data splitting. *WDSs 10 proceedings of contributed papers, Part*, 1:31–36, 2010.

[77] Paulo J Ribeiro Jr and Peter J Diggle. geor: a package for geostatistical analysis. *R news*, 1(2):14–18, 2001.

[78] Lloyd L Schulman and Joseph S Scire. Buoyant line and point source (blp) dispersion model user's guide. final report. Technical report, Environmental Research and Technology, Inc., Concord, MA (USA), 1980.

[79] Lloyd L Schulman, David G Strimaitis, and Joseph S Scire. Development and evaluation of the prime plume rise and building downwash model. *Journal of the Air & Waste Management Association*, 50(3):378–390, 2000.

[80] Joseph S Scire, Francoise R Robe, Mark E Fernau, and Robert J Yamartino. A user's guide for the calmet meteorological model. *Earth Tech, USA*, 37, 2000.

[81] Joseph S Scire, David G Strimaitis, Robert J Yamartino, et al. A user's guide for the calpuff dispersion model. *Earth Tech, Inc. Concord, MA*, 2000.

[82] EMC Education Services. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, 2015.

[83] Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, ACM '68, pages 517–524, New York, NY, USA, 1968. ACM.

[84] BP Shumaker and RW Sinnott. Astronomical computing: 1. computing under the open sky. 2. virtues of the haversine. *Sky and telescope*, 68:158–159, 1984.

[85] Kunwar P Singh, Shikha Gupta, and Premanjali Rai. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment*, 80:426–437, 2013.

[86] Jeanne Mager Stellman et al. *Enciclopedia de salud y seguridad en el trabajo*. Ministerio de Trabajo y Asuntos Sociales, Subdirección General de Publicaciones, 1999.

[87] Jason G. Su, Michael Brauer, Bruce Ainslie, Douw Steyn, Timothy Larson, and Michael Buzzelli. An innovative land use regression model incorporating meteorology for exposure analysis. *Science of The Total Environment*, 390(2–3):520 – 529, 2008.

[88] The European Environment Agency . Soer 2015 — the european environment — state and outlook 2015. http://www.eea.europa.eu/soer, 2015.

[89] Sandhya Vijayasarathy and Jhinuk Chatterjee. Comparison of mlr, isotonic regression and knn based qsar models for the prediction of inhibitory activity of hdac6 inhibitors. *International Journal of Life Sciences Biotechnology and Pharma Research*, 4(2):127, 2015.

[90] Elissa H. Wilker, Sarah R. Preis, Alexa S. Beiser, Philip A. Wolf, Rhoda Au, Itai Kloog, Wenyuan Li, Joel Schwartz, Petros Koutrakis, Charles DeCarli, Sudha Seshadri, and Murray A. Mittleman. Long-Term Exposure to Fine Particulate Matter, Residential Proximity to Major Roads and Measures of Brain Structure. *Stroke*, April 2015.

[91] Simon Wood and Maintainer Simon Wood. The mgcv package. *www. r-project. org*, 2007.

[92] World Health Organisation. Public health, environmental and social determinants of health. http://www.who.int/phe/health_topics/outdoorair/databases/health_impacts/en/, 2015.

[93] Yihui Xie. animation: An r package for creating animations and demonstrating statistical methods. *Journal of Statistical Software*, 53(1):1–27, 2013.

[94] Junsub Yi and Victor R Prybutok. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution*, 92(3):349–357, 1996.

Este trabajo ha sido financiado parcialmente por el proyecto REFRAME, subvencionado por the European Coordinated Research on Long-term Challenges in Information and Communication Sciences Technologies ERA-Net (CHIST-ERA) financiado por MINECO en España (PCIN-2013-037) y por the French National Research Agency (ANR) en Francia; y por la Generalitat Valenciana con la ayuda PROMETEOII/2015/013 (SmartLogic).

APÉNDICE A

Publicaciones relacionadas con el trabajo (Texto completo)

A.1 *Wind-sensitive Interpolation of Urban Air Pollution Forecasts*

Lidia Contreras and Cèsar Ferri

International Conference on Computational Science, California (USA), 2016 [31]

Procedia Computer Science

Volume 80, 2016, Pages 313–323

ICCS 2016. The International Conference on Computational Science

Wind-sensitive Interpolation of Urban Air Pollution Forecasts

Lidia Contreras and Cèsar Ferri

DSIC, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain.
liconoc@upv.es,cferri@dsic.upv.es

Abstract

People living in urban areas are exposed to outdoor air pollution. Air contamination is linked to numerous premature and pre-native deaths each year. Urban air pollution is estimated to cost approximately 2% of GDP in developed countries and 5% in developing countries. Some works reckon that vehicle emissions produce over 90% of air pollution in cities in these countries. This paper presents some results in predicting and interpolating real-time urban air pollution forecasts for the city of Valencia in Spain. Although many cities provide air quality data, in many cases, this information is presented with significant delays (three hours for the city of Valencia) and it is limited to the area where the measurement stations are located. We compare several regression models able to predict the levels of four different pollutants (NO, NO_2, SO_2, O_3) in six different locations of the city. Wind strength and direction is a key feature in the propagation of pollutants around the city, in this sense we study different techniques to incorporate this factor in the regression models. Finally, we also analyse how to interpolate forecasts all around the city. Here, we propose an interpolation method that takes wind direction into account. We compare this proposal with respect to well-known interpolation methods. By using these contamination estimates, we are able to generate a real-time pollution map of the city of Valencia.

Keywords: Machine Learning, Urban Air Pollution, Spatial Interpolation

1 Introduction

Air pollution is one of the factors with major impact on the health of people. Exposure to ambient air pollution increases the risk of suffering respiratory diseases, such as pneumonia, or chronic, such as lung cancer or cardiovascular diseases [21]. A recent work [20] relates structural changes in the brain to long-term exposure to ambient air pollution.The SOER 2015 report [19] concludes that although the atmosphere in Europe has improved in the last decades, there are significant traces of the most harmful contaminants. The report estimates that in 2011, 430.000 Europeans died prematurely because of pollution. In this context, citizens of urban

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.343

agglomerations must try to reduce their exposition to urban air pollution as much as possible. This is especially relevant for high risk population such as: kids, elderly people, asthmatics or people suffering respiratory diseases.

In this work we study the prediction of urban air pollution in real-time by employing historical data. We concentrate on four pollutants (NO, NO_2, O_3, SO_2). For that reason we employ data from the city of Valencia in Spain. Valencia is a medium size urban agglomeration (around 1.000.000 inhabitants). The city provides an open data site with information about: traffic data, noise levels and air pollution... Data about pollutant levels need to be verified and it is published with a delay of three hours. This delay can represent a problem since risky high levels of pollution are not detected in real-time. Moreover, the network of sensors is limited (six in the city of Valencia). Considering these restrictions, we address the problem of producing real-time predictions of the levels of pollution all around the city. We will study the performance of the predictions of different techniques for building regression models that are trained using features that represent traffic intensity, persistence of pollutants and meteorological parameters. We also analyse how the direction of wind affects the level of pollution and how to use that information in order to increase the accuracy of the prediction models.

Additionally, we address how to interpolate predictions and, in this way, we are able to show the approximate concentration of pollutants all around the city. For that reason we analyse popular spatial interpolation methods [10] such as Inverse Weighting Distance (IDW) or Kriging. These methods are static in the sense that they do not consider context conditions of the points to interpolate. Meteorological parameters (specially wind condition) clearly affects the way in which the pollutants are dispersed around the city. We propose a new method that uses wind information in order to improve the interpolation of urban pollution. In this aspect, we consider this technique a wind-sensitive interpolation approach. Experiments using a cross validation methodology show that this new method get better forecasts in comparison with well-known methods, specially when the interpolation is computed using enough information. This paper can be considered an extension of [13]. That work was focused on presenting the *Airvlc* application for real-time forecasts of air pollutants. In the models of [13] we did not consider wind direction in the learning models and interpolation techniques were not studied.

The paper is organised as follows. Section 2 details the process of data collection of pollution particles and some factors that affect the generation or dispersion of these pollutants. We also include some experiments in learning regression models for predicting the pollutant concentrations and some results on including wind direction in the models. We study some interpolations methods in Section 3. Finally, Section 4 closes the paper with a discussion of the main conclusions and some plans for future work.

2 Prediction of urban pollution

2.1 Data collection

The historical pollution data for this work has been obtained from the open data web of the Generalitat Valenciana[1]. The following particles are studied in this work:

- **NO (Nitrogen monoxide):** Nitrogen monoxide is a highly unstable compound; it causes nitrogen dioxide by quickly reacting in the atmosphere. This instability makes the nitrogen monoxide a radical whose effects on the body are abnormal DNA, lipids and proteins.

[1] http://www.cma.gva.es/cidam/emedio/atmosfera/jsp/historicos.jsp

This kind of changes derives in the medium and long term as a greater chance of developing cancer.Its origin stems largely from vehicle engines.

- **NO_2 (Nitrogen dioxide):** Nitrogen dioxide is not a directly generated pollutant, since its presence in the atmosphere is caused by the oxidation of nitrogen monoxide. In the presence of moisture, this compound results in nitric acid, and its inhalation, even in low concentrations, can cause lung tissue degradation, as well as can reduce the efficacy of the immune system, especially in children.
- **SO_2 (Sulphur dioxide):** It is a toxic gas primarily produced for sulphuric acid manufacture. SO_2 emissions are related to acid rain and atmospheric particulates. Inhaling SO_2 is associated with respiratory diseases and premature death.
- **O_3 (Ozone):** Ozone is not released directly by specific sources. This pollutant is generated by sunlight acting on NOx and Volatile organic compounds (VOC) in the air. Exposure to ozone significantly reduces lung function and induces respiratory inflammation. It can also produce symptoms such as chest pain, coughing, and pulmonary congestion.

The main sources of pollution in developed countries are motor vehicles and industry. It is useful to measure the level of traffic in a city in order to predict air pollution. The City of Valencia provides a network of sensors (electromagnetic coils) that measure the intensity of traffic (Vehicles/hour). This information can be found in the open data site of the Valencia City Council[2]. Meteorological conditions influence severely in the generation and distribution of air pollutants. In an ordinary atmosphere situation, temperature decreases with altitude, favouring ascension of warmer (and less dense) air, and dragging contaminants upwards. In a situation of thermal inversion, a warmer layer of air is over the colder surface air and prevents the rise of this last (denser), so the contamination is confined and increases. Strong winds can disperse pollutants and transport them away from their emission point. We have collected Meteorological observations of Valencia city from Meteorological Agency of the Government of Spain (AEMET)[3].

With all the selected parameters, we have built datasets aimed to predict the concentration of pollutants from the intensity of traffic and weather parameters. Concretely, we have collected data for a period of two years (2013 and 2014). Data was collected every 60 minutes, 24 hours a day during those two years. Valencia city has six stations for the detection and measurement of air pollution, although not all the stations measure the same parameters. For each one of these stations, we create a dataset with the level of the pollutants measured and parameters that can affect these measurements, we concentrate on traffic level calendar features and weather conditions. Concretely, we extract the following set of features for each station:

- **Meteorological conditions:** Temperature (Celsius degrees), Relative humidity (Percentage), Pressure (hPa), Wind speed (m/s), Rain (mm/h)
- **Calendar features:** Year, Month, Day in the month, Day in the week, Hour
- **Traffic intensity features:** Traffic level in the surrounding stations (vehicles/hour) and traffic level 3 hours before
- **Pollution features:** Pollution level in the target station 3 hours before

[2] http://www.valencia.es/ayuntamiento/DatosAbiertos.nsf/
[3] http://www.aemet.es/

Additionally, given the particular behaviour of the set of pollutants analysed and in light of that some of these pollutants can derive from others in some cases we add extra features. Precisely, for predicting NO_2 we include NO and SO_2 3 hours before, for predicting NO we also use NO_2 and SO_2 3 hours before, and finally, for forecasting O_3 we include NO and NO_2.

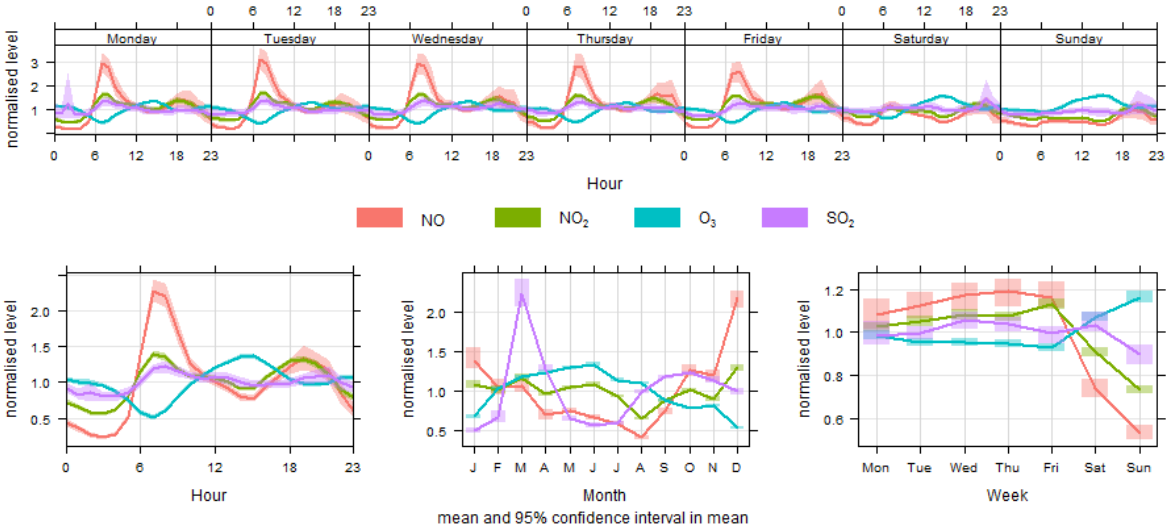

Figure 1: Distribution of the average level of four pollutants (NO, NO_2, O_3, SO_2) in *Pista* station depending on hour of the week (top), hour of the day (bottom left), month (bottom centre), week day (bottom right).

We can see a summary of the datasets in Table 1. Averages and standard deviation for the analysed stations of the pollutant particles measured and the intensity of traffic associated with each station (average traffic per hour measured the traffic sensors closer than 1km to the station) are included in this table. If we analyse traffic intensity, *Pista* and *Viveros* are the busiest stations. With regard to pollution levels *Pista* station presents the maximum levels for all the parameters except O_3. This behaviour can probably be associated with the specific situation of the station. *Pista* is located in a the central part of the city surrounded by busy streets, and therefore vulnerable to the overall city pollution. Note that some stations do not measure all the pollutants.

Figure 1 represents the distribution of the average level of four pollutants (NO, NO_2, O_3, SO_2) depending on different calendar factors[4]. It is easy to see that there are direct correspondence between the level of measured pollution and some of these factors. For instance, most of pollutants reach the lowest levels during weekend days and summer months when traffic is not intense. We can observe a peak in March in SO_2, this is a local phenomenon caused by the *fallas* traditional celebration that concludes around midnight on March 19th with the combustion of hundreds of cardboard monuments. We can also see a negative correlation between O_3 and the other three pollutants (specially NO, NO_2). This can be explained if we consider that part of the urban O_3 generation occurs when nitrogen oxides (NOx) and other compounds react in the atmosphere in the presence of sunlight. This effect can be observed in the high levels of ozone

[4]This figure has been generated using the R Openair Library [2].

around midday and in summer. The strange behaviour of high values in urban O_3 has been detected in several cities. In [3], the authors defend that "the primary cause of the higher O_3 on weekends is the reduction in oxides of nitrogen (NOx) emissions on weekends in a volatile organic compound (VOC)-limited chemical regime".

Station	Traffic		NO		NO_2		O_3		SO_2	
	ave	sd	ave	sd	ave	sd	ave	sd	ave	sd
Moli	13357	10302	10.3	19.9	28.2	20.7	46.9	25.5	2.4	2.1
Pista	34606	24462	23.5	33.8	45.1	27.0	47.1	25.0	3.8	3.8
Francia	20037	14277	8.5	18.8	26.7	23.2	50.2	25.2	2.3	2.3
Viveros	33214	24746	9.7	21.5	29.4	24.2	45.2	28.5	2.7	2.5
Bulevar	11352	8555	12.8	27.5	29.2	22.1	48.5	28.3	2.2	2.1
UPV	11987	8938	7.6	17.9	24.3	24.1	56.3	27.3	2.2	3.1

Table 1: Averages and standard deviation of the pollution detection sensors for the available stations.

2.2 Experiments

We use several regression learning techniques from R [14] in order to identify the technique that is able to better predict the levels of pollution. We build the models using as training data the registers of 2013 and the first nine months of 2014, and test the models with the last three months of 2014. Concretely, we employ the following techniques for learning regression models (all of them with the default parameters, unless stated otherwise): Linear Regression (*lr*) [5], quantile regression (*qr*) [9] with *lasso* method, K nearest neighbours (*IBKreg*) with $k = 10$ [5] , a decision tree for regression (*M5P*) [5], and Random Forest (*RF*) [11]. In order to compare the predictive performance of the regression models, we introduce three baseline models: A model that always predicts the mean of the train data (*TrainMean*), a model that always predicts the mean of the test data (*TestMean*), and a basic model that predicts the same value of the target pollutant 3 hours before (*X3H*). Root Mean Squared Error (RMSE) is used as performance measure.

Table 2 contains the RMSE of the regression models for the prediction of the four target pollution levels of the *Molí* station[5]. When observing these results, we can conclude that machine learning models are able to improve the performance of the basic baseline models in almost all cases. Comparing learning techniques, ensembles of decision trees technique (Random Forest) is the best model in almost all of cases. Given these results, Random Forest models will be applied in the following experiments in this work.

Machine learning has been widely used for predicting pollution levels. A seminal work in this area with neural networks is [22]. Neural networks models have been widely employed in this field, a review of these approaches can be found in [8]. A more related work is [7]. Here the authors propose a modelling system for predicting the traffic volumes, emissions from stationary and vehicular sources, and atmospheric dispersion of pollution in an urban area. The paper compares the predicted NO and NO_2 concentrations with the results of an urban air quality monitoring network. The agreement of model predictions was better for the two suburban monitoring stations, compared with two urban stations. Our comparison of regression techniques obtains similar conclusions to the work presented in [17]. In this study, principal components analysis (PCA) is performed to identify air pollution sources. From the extracted

[5] Results for the other five stations are shown in http://www.dsic.upv.es/~flip/pollutionexp.pdf.

features, tree based ensemble learning models are induced to predict the urban air quality of Lucknow (India) together with the air quality and meteorological databases for a period of five years.

	TrainMean	TestMean	X3h	LR	qr	IBkreg	M5p	RF
NO	30.41	28.45	33.76	25.32	30.29	27.33	23.55	**19.72**
NO2	20.80	20.80	20.72	15.52	15.47	15.04	16.51	**14.42**
O3	30.33	21.81	18.32	14.91	14.98	15.88	12.52	**11.79**
SO2	1.65	1.14	1.25	1.13	1.25	1.26	1.18	**1.04**

Table 2: Results in RMSE of different regression models for Moli Station. The best prediction model is highlighted in bold.

2.2.1 Models with wind direction

Previous methods only take wind strength into account in order to build machine learning models. Wind direction can contribute significantly in the dispersion of pollutants. For instance, if we consider quarters in the shore of coastal cities, when wind is coming from sea, the levels of pollutants are drastically lower than when winds is coming from dense populated areas of the city. This behaviour can be seen in Figure 2 (generated with library [2]). These plots show the average of NO and NO_2 pollutants depending on wind speed and direction, and they clearly show how these factors correlate with respect to the level of these pollutants.

A simple way of using wind direction component is to consider the pair attributes sine and cosine of the angle defined by wind direction. In our case, this method has obtained poor results. Therefore we have adopted a different approach of including wind direction: we modify the area where we select sensors for the traffic measurement according to wind direction. We use the idea that traffic is generating many of the pollutants that we are trying to predict, and these pollutants are dispersed according to wind speed and direction. We study two different versions of this idea: in the *dir* method we consider the traffic that is generated in the radius of 1km from the sensor but only considering the traffic sensors that are in the windward circular sector of 30º; and in the *wdir* method is similar to *dir* method but now we use the windward in order to weight traffic sensors, and in this way we give more importance to traffic measures in the circular sector of 30º. In Table 3 we compare the results in RMSE of the six stations (with the same methodology of the previous section) using three scenarios to incorporate wind direction: *nd* (wind direction is not used), *dir* method and *wdir* method. The results show that performance of the methods depends drastically on the pollutant. *dir* and, specially, *wdir* methods are able to improve the prediction performance in particles directly related to traffic emissions (SO_2, NO and NO_2) probably because these techniques of modelling wind direction are based on the selection of traffic measures according to the direction of the wind stream. O_3 is not directly related to traffic emissions, and in this case, the *dir* and *wdir* methods are generally not able to enhance the *nd* baseline method. In any case we can also observe a wide variety of performance depending on the station since every station is located in a different environment with specific features (city centre, residential area, coast shore...).

Wind direction has rarely been incorporated into land-use regression models. [4] identifies 25 land-use regression studies and only two incorporate wind direction in the predictive models. [1] studied the use of wind fields to improve the prediction of air pollution in Toronto. Wind direction fields were constructed from 38 weather stations, and these features were significantly useful for NO_2 prediction. Another approach is [18]. Here, the authors apply land use regression

(LUR) integrating wind speed, wind direction and cloud cover/insulation to estimate hourly NO and NO_2 concentrations.

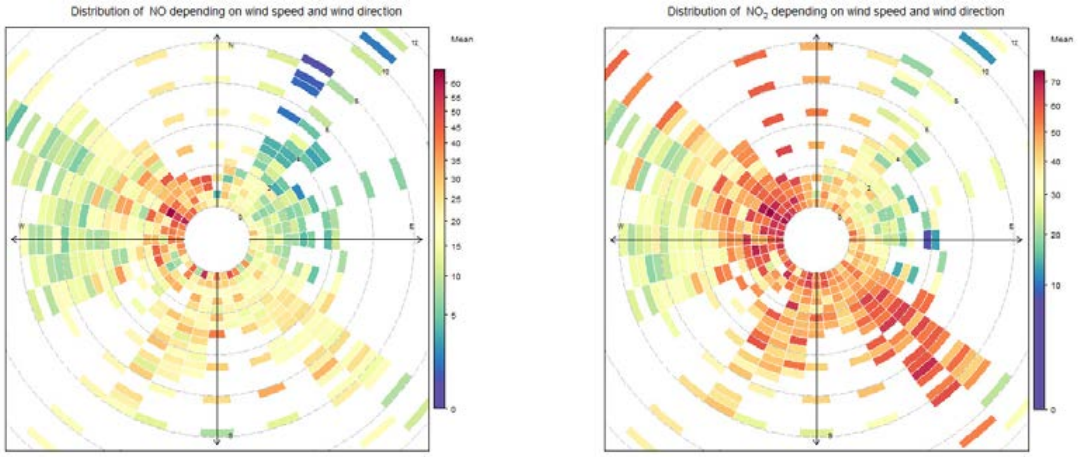

Figure 2: Distribution of NO and NO_2 pollutants in Pista station depending on wind speed and wind direction.

Station	NO			NO_2			O_3			SO_2		
	nd	*dir*	*wdir*	*nd*	*dir*	*wdir*	*nd*	*dir*	*wdir*	*nd*	*dir*	*wdir*
Moli	19.72	20.50	**19.54**	**14.42**	15.78	14.50	11.79	**11.67**	12.01	1.038	**1.037**	1.046
Pista	32.98	33.90	**32.48**	17.63	**16.73**	17.55	14.48	14.92	**14.41**	1.76	**1.74**	1.75
Francia	22.65	**23.53**	22.64	13.53	14.57	**13.44**	**12.61**	13.00	12.76	2.23	2.25	**2.22**
Viveros	26.38	26.46	**25.96**	**12.79**	13.49	12.99	13.30	13.50	**13.29**	**1.17**	1.19	1.18
Bulevar	28.79	30.21	**28.46**	**15.26**	18.18	15.77	11.37	**11.21**	11.42	0.92	1.14	**0.91**
UPV	26.21	26.32	**26.18**	18.71	19.17	**18.51**	12.84	**12.44**	12.89	0.90	0.91	**0.85**

Table 3: RMSE of the Random Forest regressors for the pollutants NO, NO_2, O_3 and SO_2 depending on the method to model wind direction: *nd* (wind direction is not used), *dir* (direction is used to select traffic sensors), *wdir* (similar to *dir* method, but nearby traffic sensors are given more relevance). The best prediction model is highlighted in bold.

3 Interpolation of predictions

In the previous section we have analysed how to obtain real-time air pollution predictions from a given set of features. Our objective in this section is to interpolate predictions all around the city in order to be able of forecasting the concentration of pollutants in locations that are not close to the pollutant measurement station. Spatial interpolation [10] tries to predict values for cells in a raster from a limited number of sample data points. Spatial interpolation can be used to forecast unknown values (eg. elevation, chemical concentrations, noise levels..) for any geographic point data in the raster. Formally, given a set of N known sample data points in the study region $\mathcal{D}$. The set of N known data points are a list of tuples: $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}, x \in \mathcal{D}, y \in \mathbb{R}$. A spatial interpolation method is a function u, such that $u(x) : x \rightarrow \mathbb{R}, x \in \mathcal{D}$. We study the following interpolation techniques:

- **Mean**: A baseline method where we always predict the average of all the N known points ($\overline{y}$). Formally, $u(x) \rightarrow \overline{y}$.

- **Inverse Distance Weighting (IDW)**: The values of unknown points are computed using a weighted average of known points. The weights are estimated using distances among the target point and known points. Here we used the well-known *Shepard's method* [16] with power parameter $p = 1$.
- **Local Inverse Distance Weighting (LIDW)**: A different method for Inverse Distance Weighting. This version assigns greater influence to values closest to the interpolated point compared to IDW and $p = 1$, We define $d(x_a, x_b)$ as the Euclidean distance in $\mathcal{D}$ between points x_a and x_b. Then $u(x_i) = w(x_1, x_i) * y_i + ..w(x_N, x_i) * y_N$, where $w(j,k) = (D_{tot}(k) - d(j,k))/(D_{tot}(k) * (|N| - 1))$, and $D_{tot}(k) = \sum(d(k, x_i)), \forall x_i \in N$.
- **Wind Sensitive LIDW**: A modification of LIDW that takes into account wind direction in such a way that we increase the weights of the known points that are windward. We define a windward circular sector of 30º from the point to predict, and the weights of the stations located in that sector are increased by a factor $\alpha = 1.5$.
- **Kriging**: In Kriging the surrounding measured values are weighted to produce a predicted value for an unknown point. Weights are based on the distance between the known points, the prediction locations, and the overall spatial arrangement among the known points. Here we use the R implementation of [15].

In order to evaluate the interpolation methods, from the six available stations, we establish three different settings. A) 3 stations as known points versus 3 stations as unknown points (20 possible combinations); B) 4 stations as known points versus 2 stations as unknown points (15 possible combinations); C) 5 stations as known points versus 1 stations as unknown points (6 possible combinations). Table 4 includes the average RMSE of the unknown points with respect the real value for the three different settings and the studied pollutants. Here, we use the whole dataset, i.e. hourly measures for years 2013 and 2014. If we compare the results of the interpolation methods in this table, Kriging and Wind Sensitive LIDW obtain the best performance. In general, Kriging interpolates better with few known points and Wind Sensitive LIDW shows better performance when it can use more information to interpolate. The exception to this behaviour is O_3 where interpolation methods cannot improve the mean baseline.

	A)**3 Known - 3 Unknown 20it**					B)**4 Known - 2 Unknown 15it**					C)**5 Known - 1 Unknown 5it**				
	Mean	LIDW	IDW	Wind	Krig.	Mean	LIDW	IDW	Wind	Krig.	Mean	LIDW	IDW	Wind	Krig.
NO	22.15	19.55	17.67	**16.92**	19.94	18.13	17.76	17.89	**14.99**	19.89	17.13	16.72	14.76	**14.71**	22.98
NO_2	21.49	22.06	21.46	20.02	**19.99**	19.04	22.19	20.80	**17.31**	20.56	17.90	19.20	16.31	**16.09**	24.18
O_3	15.37	14.54	17.54	17.72	**13.16**	**12.32**	20.44	14.44	15.13	12.66	**12.03**	13.74	13.87	13.90	18.66
SO_2	3.05	3.23	2.95	2.91	**2.90**	3.19	2.96	3.05	**2.47**	3.42	2.97	2.90	**2.37**	2.47	3.42

Table 4: Comparison of five methods of spatial interpolation of four pollutants and three different settings. RMSE of of the unknown points with respect the actual value for the hourly measures of years 2013 and 2014. The best prediction model is highlighted in bold.

The application of spatial interpolation methods over the forecasts of the six pollution stations provides a way to estimate a real-time pollution heat map of the city. An example of these plots are included in Figure 3. Here we show the spatial interpolation of O_3 by Wind Sensitive LIDW (left) and Kriging (right) in the city of Valencia.

Some works have addressed the interpolation of air pollution forecasts. In [12] the authors compare Land-use regression (LUR) [4] and universal Kriging (UK) (a version of Kriging that assumes a general polynomial trend model). In their experiments with prediction models for NOx in Los Angeles (USA), the UK interpolation consistently outperformed LUR. The RIO method is presented in [6] as a interpolation model for air pollution. The method uses a β

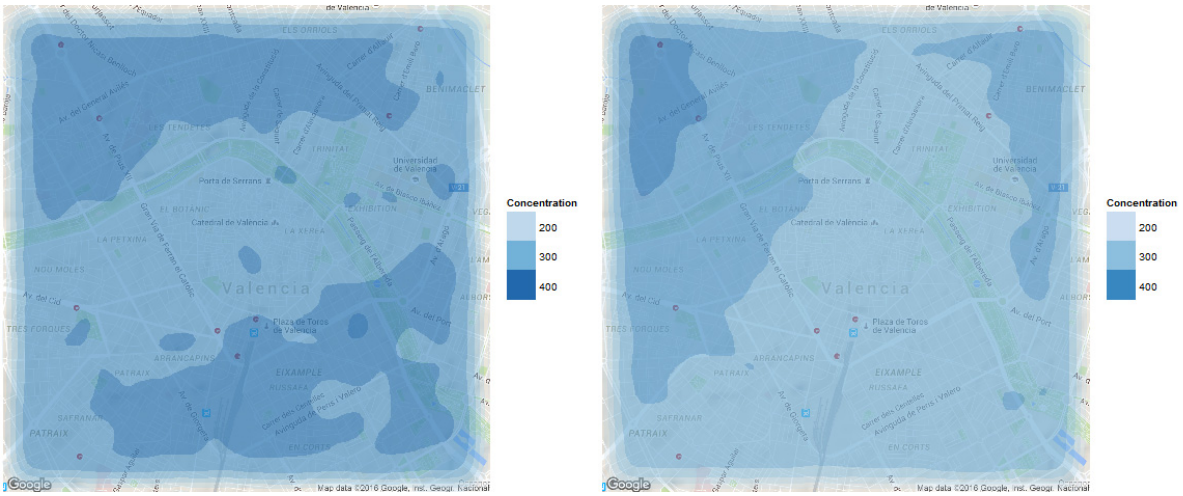

Figure 3: Spatial interpolation of O_3 by Wind Sensitive LIDW (left) and Kriging (right).

parameter that offers flexibility in the weighting between land use and air pollution levels. The experiments with O_3, NO_2 and PM_{10} in a cross-validation procedure show that RIO produces better results compared to IDW and Ordinary Kriging.

4 Conclusions

Poor Air quality is one of the factors that can decrease life expectancy since contamination rises the risk of suffering respiratory diseases. The detection of risky levels in real time can reduce the exposure to ambient air pollution. In this work, we have studied machine learning methods that predicts in real-time the levels of four dangerous pollutants for the six pollution measurement stations in the city of Valencia. According to our experiments, the Random Forest technique is able to build the best forecast models in most of the studied cases. We have analysed how we can enrich these models by incorporating wind direction information. We have proposed an approach where wind direction is used for dynamically select the traffic emission sources. The results show that this approach is able to improve the performance of the predictions for the pollutants directly related to emissions by fuel combustion. Finally, we have proposed a new interpolation method based on LIDW (Local Inverse Distance Weighting) that takes wind direction into account. We have compared the novel technique with respect to well-known spatial interpolation methods such as Kriging or common IDW. The experiments show that our Wind Sensitive LIDW obtains a positive performance specially when there are significant number of known points to use in the interpolation.

As future work, we propose the application of local features of the target points in the interpolation methods, e.g. nearby traffic level or altitude. We also plan to apply the presented techniques in other cites in order to study if similar behaviours are observed.

Acknowledgments

This work has been partially supported by the REFRAME project, granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences Technologies ERA-Net (CHIST-ERA) funded by MINECO in Spain (PCIN-2013-037), the EU (FEDER) and the Spanish MINECO under grants TIN 2015-69175-C4-1-R and TIN 2013-45732-C4-1-P and by Generalitat Valenciana under grant PROMETEOII/2015/013. We are also grateful to Ajuntament de València, InnDEA-València and specially to Ramón Ferri and Ruth López for their help in providing traffic data.

References

[1] M.A. Arain, R. Blair, N. Finkelstein, J.R. Brook, T. Sahsuvaroglu, B. Beckerman, L. Zhang, and M. Jerrett. The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. *Atmospheric Environment*, 41(16):3453 – 3464, 2007.

[2] David C. Carslaw and Karl Ropkins. openair — an r package for air quality data analysis. *Environmental Modelling and Software*, 27–28(0):52–61, 2012.

[3] Jon M Heuss, Dennis F Kahlbaum, and George T Wolff. Weekday/weekend ozone differences: what can we learn from them? *Journal of the Air & Waste Management Association*, 53(7):772–788, 2003.

[4] Gerard Hoek, Rob Beelen, Kees de Hoogh, Danielle Vienneau, John Gulliver, Paul Fischer, and David Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33):7561 – 7578, 2008.

[5] Kurt Hornik, Christian Buchta, and Achim Zeileis. Open-source machine learning: R meets Weka. *Computational Statistics*, 24(2):225–232, 2009.

[6] Stijn Janssen, Gerwin Dumont, Frans Fierens, and Clemens Mensink. Spatial interpolation of air pollution measurements using {CORINE} land cover data. *Atmospheric Environment*, 42(20):4884 – 4903, 2008.

[7] A Karppinen, J Kukkonen, T Elolähde, M Konttinen, and T Koskentalo. A modelling system for predicting urban air pollution:: comparison of model predictions with the data of an urban measurement network in helsinki. *Atmospheric Environment*, 34(22):3735–3743, 2000.

[8] Mukesh Khare and SM Shiva Nagendra. *Artificial neural networks in vehicular pollution modelling*, volume 41. Springer, 2006.

[9] Roger Koenker. *quantreg: Quantile Regression*, 2015. R package version 5.11.

[10] Jin Li and Andrew D Heap. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. *Ecological Informatics*, 6(3):228–241, 2011.

[11] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[12] Laina D Mercer, Adam A Szpiro, et al. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (nox) for the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Atmospheric Environment*, 45(26):4412–4420, 2011.

[13] Lidia Contreras Ochando, Cristina I. Font Julián, Francisco Contreras Ochando, and Cèsar Ferri Ramirez. Airvlc: An application for real-time forecasting urban air pollution. In *Proceedings of the 2nd International Workshop on Mining Urban Data*, pages 72–79, 2015.

[14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[15] P.J. Ribeiro Jr. and P.J. Diggle. geoR: a package for geostatistical analysis. *R-NEWS*, 1(2):15–18, 2001.

[16] Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, ACM '68, pages 517–524, New York, NY, USA, 1968. ACM.

[17] Kunwar P Singh, Shikha Gupta, and Premanjali Rai. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment*, 80:426–437, 2013.

[18] Jason G. Su, Michael Brauer, Bruce Ainslie, Douw Steyn, Timothy Larson, and Michael Buzzelli. An innovative land use regression model incorporating meteorology for exposure analysis. *Science of The Total Environment*, 390(2–3):520 – 529, 2008.

[19] The European Environment Agency . Soer 2015 — the european environment — state and outlook 2015. http://www.eea.europa.eu/soer, 2015.

[20] Elissa H. Wilker, Sarah R. Preis, Alexa S. Beiser, Philip A. Wolf, Rhoda Au, Itai Kloog, Wenyuan Li, Joel Schwartz, Petros Koutrakis, Charles DeCarli, Sudha Seshadri, and Murray A. Mittleman. Long-Term Exposure to Fine Particulate Matter, Residential Proximity to Major Roads and Measures of Brain Structure. *Stroke*, April 2015.

[21] World Health Organisation. Public health, environmental and social determinants of health. http://www.who.int/phe/health_topics/outdoorair/databases/health_impacts/en/, 2015.

[22] Junsub Yi and Victor R Prybutok. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution*, 92(3):349–357, 1996.

A.2 *Airvlc: An application for real-time forecasting urban air pollution*

Lidia Contreras Ochando, Cristina I. Font Julián, Francisco Contreras Ochando and Cèsar Ferri

2nd International Workshop on Mining Urban Data co-located at 32th International Conference on Machine Learning, Lille (France), 2015 [68]

Airvlc: An application for real-time forecasting urban air pollution

Lidia Contreras Ochando LICONOC@UPV.ES
Universitat Politècnica de València. Spain

Cristina I. Font Julián CRIFONJU@EI.UPV.ES
Universitat Politècnica de València. Spain

Francisco Contreras Ochando FRACONOC@GMAIL.COM
Universitat Politècnica de València. Spain

Cèsar Ferri CFERRI@DSIC.UPV.ES
DSIC. Universitat Politècnica de València. Spain

Abstract

This paper presents Airvlc, an application for producing real-time urban air pollution forecasts for the city of Valencia in Spain. Although many cities provide air quality data, in many cases, this information is presented with significant delays (three hours for the city of Valencia) and it is limited to the area where the measurement stations are located. The application employs regression models able to predict the levels of four different pollutants (CO, NO, PM2.5, NO2) in three different locations of the city. These models are trained using features that represent traffic intensity, persistence of pollutants and meteorological parameters such as wind speed and temperature. We compare different learning techniques to get the better performance in the prediction of pollutants. According to our experiments, ensembles of decision trees (Random Forest) outperforms the rest of methods in almost all of our tests. Airvlc incorporates the best regression models and, by a distance-weighted combination of the predictions, is able to generate a real-time pollution map of the city of Valencia. The application also includes a warning system for sending notifications to users when a nearby risk pollution concentration is detected.

Proceedings of the 2^{nd} International Workshop on Mining Urban Data, Lille, France, 2015.

1. Introduction

Air pollution can have important impact (short and long-term) on the health of people. For instance, urban air pollution increases the risk of suffering respiratory diseases such as pneumonia, or chronic, such as lung cancer or cardiovascular disease (World Health Organisation, 2015). A recent work (Wilker et al., 2015) relates long-term exposure to ambient air pollution to structural changes in the brain.The SOER 2015 report (The European Environment Agency , 2015), with data about the European Union countries' air quality in 2011, concludes that although the atmosphere in the continent has improved in the last decades, there are significant traces of the most harmful contaminants. In fact, in 2011, the report estimates that 430.000 Europeans died prematurely because of pollution.

Although some governments are introducing restriction policies that limit the use of vehicles (main source of pollution in most cases), only in Europe, important cities such as Paris, Naples, Moscow, Milan or Barcelona still report significant levels of urban pollution in 2015 (The European Environment Agency , 2015). In this context, it is important for citizens of urban agglomerations to reduce the exposition to urban air pollution as much as possible. This is especially relevant for high risk population such as: kids, elderly people, asthmatics or people suffering respiratory diseases.

In this work we present an application that predicts urban air pollution in real time by employing historical data. The application is based on the city of Valencia in Spain. This city can be considered a medium size urban agglomeration (around 1.000.000 inhabitants). The city provides an open data site containing real-time information about the city in different aspects such as traffic data, noise sensors, pollen

sensors... Although different sensors of urban pollution air are included in the site, this information needs to be carefully verified and it is published with a delay of three hours. This delay can represent a problem since risky high levels of pollutions are not detected in real-time. Additionally, the network of sensors is limited (six in the city of Valencia).

Considering these limitations, we have developed an application able to display in real-time foreseeable levels of pollution in a wide number of points of the city. The application is based on the predictions of regression models that are trained using features that represent traffic intensity, persistence of pollutants and meteorological parameters.

The paper is organised as follows. Section 2 details the process of data recollection of pollution particles and the factors that affect the generation, concentration or dispersion of these pollutants. Experiments in learning regression models for predicting the pollutant concentrations are included in Section 3. The Airvlc application is detailed in 4. Related works are discussed in Section 5. Finally, Section 6 closes the paper with a discussion of the main conclusions and some plans for future work.

2. Data collection

Different particles are associated with urban air pollution. In order to measure air contamination, pollutant parameters found in the lower levels of the troposphere are controlled. Air quality sensors measure concentrations of particles that have an anthropogenic origin and produce effects during or after the inhalation by humans. The historical pollution data for this work has been obtained from the open data web of the Generalitat Valenciana[1]. Following the recommendations of (The European Environment Agency , 2015), we concentrate on the following particles:

- **PM 2.5 (Suspended particles below 2.5 microns):** This parameter has been chosen because of its pollutant power. It is one of the most dangerous particles, since its size makes it almost unstoppable by the natural filters of the body. This fact means that the PM 2.5 are usually able to reach the pulmonary alveoli and in some cases, these particles are attached to these alveoli with a consequent reduction of lung capacity; in worst cases, the particles cross the alveolar membranes and reach the blood stream. Considering that PM 2.5 particles have its origin in anthropogenic activities (especially in the use of fuels in motor vehicles), it is not surprising that its atomic structure contains heavy metals, extremely toxic to the human body. Atmospheric conditions in the Mediterranean coast of Spain can influence the particle levels, due to lower rainfall and wind action with respect to other northern Europe countries, and the North African particles (Saharan dust), PM10 and PM2.5.

- **NO (Nitrogen monoxide):** Nitrogen monoxide is a highly unstable compound; it causes nitrogen dioxide by quickly reacting in the atmosphere. This instability makes the nitrogen monoxide a radical, namely, a high reactive power molecule, whose effects on the body are abnormal DNA, lipids and proteins. This kind of changes derives in the medium and long term as a greater chance of developing cancer.Its origin stems largely from vehicle engines.

- **NO2 (Nitrogen dioxide):** Nitrogen dioxide is not a directly generated pollutant, since its presence in the atmosphere is caused by the oxidation of nitrogen monoxide. In the presence of moisture, this compound results in nitric acid, and its inhalation, even in low concentrations, can cause lung tissue degradation, as well as can reduce the efficacy of the immune system, especially in children.

- **CO (Carbon monoxide):** Carbon monoxide is a primary pollutant. CO is toxic; it prevents oxygen transport by poisoning the blood, since it replaces the haemoglobin. People with cardiovascular and cerebrovascular problems could suffer heart attacks or strokes because of problems related to high concentrations of CO.

The distribution of air pollution is decisively influenced by climatic conditions. We have collected Climatological observations for the meteorological data of Valencia city from Meteorological Agency of the Government of Spain (AEMET)[2]. We consider the following parameters:

- **Temperature:** In an ordinary atmosphere situation, temperature decreases with altitude, favouring ascension of warmer (and less dense) air, and dragging contaminants upwards. In a situation of thermal inversion, a warmer layer of air is over the colder surface air and prevents the rise of this last (denser), so the contamination is confined and increases.

- **Humidity:** Humidity is a weather factor to be considered; in its presence, nitrogen dioxide derives in nitric acid, harmful to human health.

- **Wind speed:** Strong winds can disperse pollutants and transport them away from their emission point.

[1] http://www.cma.gva.es/cidam/emedio/atmosfera/jsp/historicos.jsp

[2] http://www.aemet.es/

- **Precipitations** Precipitations wash contaminants and can dissolve substances and gases.

The two main sources of pollution in developed countries are motor vehicles and industry. Vehicles release large amounts of nitrogen oxides, carbon oxides, hydrocarbons and particulates when burning gasoline and diesel. Therefore, we need to measure the level of traffic in the city in order to predict the air pollution. For this purpose, the City of Valencia provides a network of sensors (electromagnetic coils) that measure the intensity of traffic (Vehicles/hour) in the city. This data can be found in the open data site of the Valencia City Council[3].

3. Experiments

With all the selected parameters, we have built datasets aimed to predict the concentration of pollutants from the intensity of traffic and weather parameters. Concretely, we have collected data for a period of two years (2013 and 2014). Data was collected every 60 minutes, 24 hours a day during those two years. Although Valencia city has six stations for the detection and measurement of air pollution, three of them have not sufficient data for the analysed period and were discarded. In this way we collected data from these stations: *Molì*, *Avd Francia* and *Pista de Silla*. These three stations are located inside the urban agglomeration, and thus most of the pollutants measured in the sensors should be generated by urban activities (mainly traffic). For each one of these stations, we create a dataset with the level of the pollutants measured and parameters that can affect these measurements, we concentrate on traffic level (measured by electromagnetic coils), weather conditions. In order to measure the traffic related to each air pollution station, we average the traffic intensity of the closest six traffic measurament sensors. This is a simplification since, certainly, all the traffic of the city has effect on the measured level of all the stations in the city.

We can see a summary of the three datasets in Table 1. This table includes averages and standard deviation for the three stations of the pollutant particles measured and the intensity of traffic associated with each station. If we analyse traffic intensity, *Avd Francia* is the busiest station, while the other two have similar values. With regard to pollution levels *Pista de Silla* station presents the maximum levels for three parameters. The only exception is PM2.5. This behaviour can probably be associated with the specific location of the stations: While *Pista de Silla* station is located in a the central part of the city, and therefore more vulnerable to the overall city pollution, the other two are in the suburbs of the city where external air streams can reduce the levels of pollutants.

We first study the weekly evolution of pollutants in the three stations. Figure 1 shows the evolution of the average of the four parameters of pollution analysed and the average traffic intensity for *Molì* station depending on the day of the week. Figure 2 presents the same plot for *Avd Francia* station and Figure 3 corresponds to *Pista de Silla* station. In order to make the values comparable in the plot we normalise each parameter by the maximum value of that parameter. The level of pollutants and traffic reach the maximum levels during the working days of the week for the three stations (Friday seems to be the worst day). We can clearly see the dependency of the four parameters of pollution on the traffic intensity level. During the weekend days, the level of traffic drastically descends and associated with this reduction the levels of pollutants significantly drop. Again, the exception is PM2.5. This behaviour can be caused because these particles can be generated by all types of combustion activities (motor vehicles, power plants, wood burning, etc.) and certain industrial processes (US Environmental Protection Agency , 2015).

We have performed a similar analysis considering the evolution of pollutants, traffic intensity and meteorological variables during a day (humidity and wind). Figure 4 shows the evolution of the daily average of these parameters for *Molì* station depending on the hour of the day . Figure 5 corresponds to *Avd Francia* station and Figure 6 to *Pista de Silla* station. Again, we normalise each parameter by the maximum value of that parameter. If we observe traffic intensity, we can discover in all the three plots a similar behaviour, there are three peaks in traffic intensity corresponding to the hours where workers travel to their work places (around 9 am), lunch time (around 2 pm) and an evening period (around 8 pm). In the three stations the maximum of pollution parameters is found at the same period of the first peak in traffic intensity (around 9 am). In the second peak of traffic intensity (around 2 pm) the levels of pollutants does not follow the increase in traffic. In fact, after the maximum period around 9 am, pollutants decrease their levels until around 4 pm where they change the behaviour and start an increasing of the values. The second peak in pollutant values is found around 9 pm. Our intuition with respect to this behaviour is that wind disperses part of the pollutant in the most sunny hours. Valencia is in the Mediterranean coast and in this city it is easy to find (especially in summer) sea breezes. These kind of winds are created over bodies of water (usually sea or big lakes) near land due to differences in air pressure created by their different heat capacity. This phenomenon can be detected in the plots if we observe the increase in wind strength during the midday hours. Finally, we observe a strange and different behaviour of the CO particle in *Molí* station. For this pollutant there is a second peak in the midday period.

[3] http://www.valencia.es/ayuntamiento/DatosAbiertos.nsf/

This behaviour probably corresponds to an extra source of pollution that needs to be further studied.

As stated previously, we are interested in predicting pollution levels in real time. Since these levels are only made public with a delay of three hours, we need to produce a prediction model from real time features. We extract the following set of features from the data collected from different sources (detailed in the previous section):

- **Climatological features:** Temperature (Celsius degrees), Relative humidity (Percentage), Pressure (hPa), Wind speed (km/h), Rain (mm/h)
- **Calendar features:** Year, Month, Day in the month, Day in the week, Hour
- **Traffic intensity features:** Traffic level in the surrounding stations (vehicles/hour), traffic level 1, 2, 3 and 24 hours before
- **Pollution features:** Pollution level in the target station 3 and 24 hours before

With this goal we compare several regression learning techniques from R (R Core Team, 2015) in order to identify the technique that is able to better predict the levels of pollution. To test the prediction ability of different models, we learn the models using as training data the registers of 2013 and the first nine months of 2014. We test the models with the last three months of 2014. We use Mean Squared Error (MSE) as a performance measure. Concretely, we employ the following techniques for learning regression models (all of them with the default parameters, unless stated otherwise): Linear Regression (*lr*) (Hornik et al., 2009), quantile regression (*qr*) (Koenker, 2015) with *lasso* method, K nearest neighbours (*IBKreg*) with $k = 10$ (Hornik et al., 2009) , a decision tree for regression (*M5P*) (Hornik et al., 2009), Random Forest (*RF*) (Liaw & Wiener, 2002), Support Vector Machines (*SVM*) (Meyer et al., 2014) and Neural Networks (Venables & Ripley, 2002). In order to compare the predictive performance of these models, we also introduce three baseline models: A model that always predicts the mean of the train data (*TrainMean*), a model that always predicts the mean of the test data (*TestMean*), and a basic model that predicts the same value of the target pollutant 3 hours before (*X3H*).

Table 2 contains the MSE of the regression models for the prediction of the four target pollution levels of the *Molí* station. Results for *Pista de Silla* station and *Avd Francia* station are shown in Table 4 and 3 respectively. If we analyse these results, we can conclude that learned models are improving the performance of the basic baseline models in almost all cases. When we compare the learning techniques in the three tables, the ensemble of decision trees technique (random forest) is the best model in almost all of cases. These results are in concordance with (Singh et al., 2013) where ensembles of trees outperformed other approaches such as SVMs.

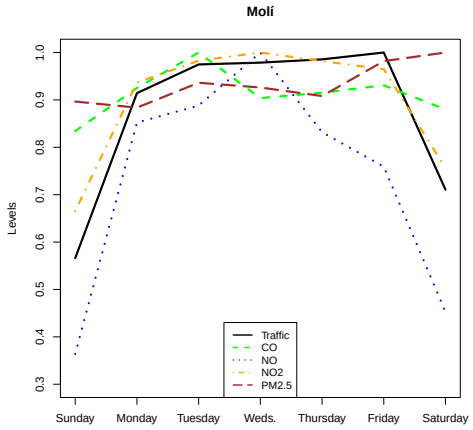

Figure 1. Average weekly traffic intensity and pollution parameters measured in Molí station.

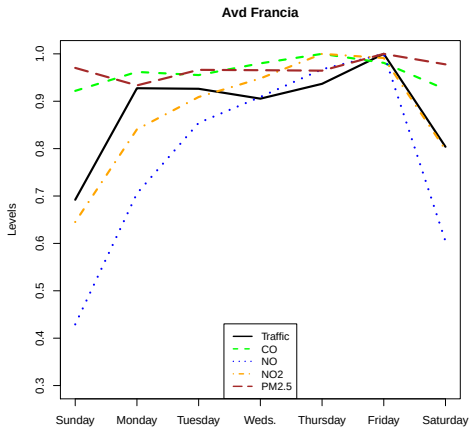

Figure 2. Average weekly traffic intensity and pollution parameters measured in Avd Francia station.

4. Airvlc

In the previous section we have analysed how to obtain real-time air pollution predictions from a given set of features. In this section we summarise Airvlc, a mobile app for Android and iOS and a web application[4]. This application generates from the regression models a map of the city of Valencia showing the predicted intensity of pollution levels. The application also allows the user to configure a set of automatic warnings every time a pollution threshold is reached near the position of the mobile device.

[4] http://airvlc.lidiacontreras.com/

Table 1. Averages and standard deviation of the three pollution detection sensors.

	Traffic		CO		NO		NO2		PM2.5	
	ave	sd	ave	sd	ave	sd	ave	sd	ave	sd
Molí	442.333	339.489	0.116	0.093	8.642	20.311	26.608	21.003	10.650	6.926
Francia	631.569	431.412	0.185	0.122	9.092	19.271	27.840	23.992	7.909	4.235
Silla	484.722	298.768	0.228	0.187	23.559	33.024	45.631	25.376	8.309	6.020

Table 2. Results in MSE of different regression models for Molí Station. The best prediction model is highlighted in bold.

	TrainMean	TestMean	X3h	lr	qr	IBkreg	M5p	RF	SVM	NN
CO	0.086	0.061	0.067	**0.057**	0.060	0.068	0.071	**0.057**	0.063	0.182
NO	30.202	28.739	36.516	25.200	29.805	27.821	25.555	**20.655**	25.870	32.944
NO2	19.918	19.914	25.258	19.680	17.370	15.683	31.242	14.877	**14.488**	32.152
PM2.5	8.803	8.803	8.634	6.889	6.564	6.674	7.248	**6.072**	6.135	13.089

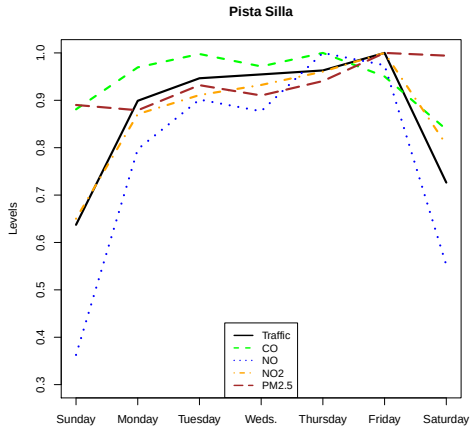

Figure 3. Average weekly traffic intensity and pollution parameters measured in Pista Silla station.

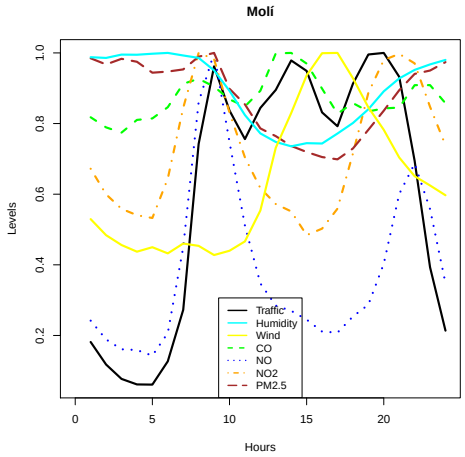

Figure 4. Average daily traffic intensity and pollution parameters measured in Molí station.

4.1. Contamination intensity map

Results of Section 3 show that random forest models obtain the best performance in most cases. Therfore, twelve random forest models are implemented in the Airvlc application. These models are able to predict every hour the level of the four analysed particles at the three pollution detection stations. We want, however, to predict pollution levels at the points of the city where the traffic is measured (1245 points around the city). For that purpose, given any of these points, we extract the features related to traffic intensity from the six nearest traffic sensors. The meteorological features are the same for all the city. The predictions of pollutants in that exact location is computed by the combination of the models corresponding to the three stations. The combination is weighted with respect to the distance of the target point with respect to the measurement stations giving more importance to the closest models. A simpler approach could be to learn a single model from the concatenation of the data from the three stations and then apply this in all the set of target points.

By computing the pollution predictions for a set of strategic and well-distributed locations we are able to estimate a real-time pollution map of the city. The map is generated with *Google Maps* technology. This map shows for each location its pollution level as a dot which colour varies among green, yellow and red depending on the calculated pollution level. If the user selects one of these dots, an extended window is opened where the exact predicted levels are shown. Figure 7 includes a screen-shot of the pollution map of the Airvlc application. The user can also select a second frame in the window of the Airvlc application where he/she can introduce a specific location and then the application computes the predicted pollution levels for that selection. An example of this process is included in Figure 8.

4.2. Risk levels

Figure 8 shows how the pollution levels are presented to users. However, showing just a concentration value of each parameter is not very useful for most users, since most of them are not experts in pollutants and they could not interpret correctly these numbers. In order to improve the comprehensibility of the predictions we have established three ranges of risk represented as speedometer: Low risk (green) corresponds to a measurement that is safe; Medium risk (yellow) when concentrations reach levels to cause harmful effects in people sensitive to air pollution exposure (kids, elderly people...); High risk (red) when concen-

Table 3. Results in MSE of different regression models for Avd. Francia Station. The best prediction model is highlighted in bold.

	TrainMean	TestMean	X3h	lr	qr	IBkreg	M5p	RF	SVM	NN
CO	0.195	0.165	0.224	0.159	0.163	0.168	0.160	**0.153**	0.156	0.324
NO	35.493	33.634	44.955	32.992	36.049	34.092	30.350	**29.517**	33.364	38.262
NO2	23.443	20.900	27.162	16.299	16.718	18.494	23.782	**14.851**	19.100	41.929
PM2.5	3.721	3.718	3.879	3.326	**3.132**	3.523	4.655	3.214	3.265	7.974

Table 4. Results in MSE of different regression models for Pista Silla Station. The best prediction model is highlighted in bold.

	TrainMean	TestMean	X3h	lr	qr	IBkreg	M5p	RF	SVM	NN
CO	0.278	0.222	0.304	0.221	0.227	0.221	0.235	**0.218**	0.268	0.278
NO	49.149	46.232	60.353	39.863	43.524	42.760	44.150	**36.332**	52.438	58.798
NO2	23.135	23.122	30.167	20.861	19.031	18.487	25.972	**16.722**	23.313	49.699
PM2.5	6.911	6.660	7.119	5.663	5.342	5.750	7.189	**5.339**	7.368	11.061

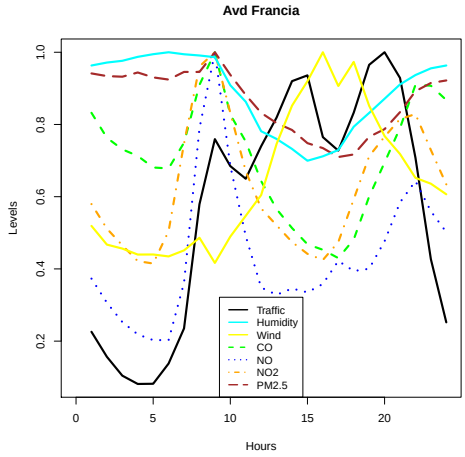

Figure 5. Average daily traffic intensity and pollution parameters measured in Avd Francia station.

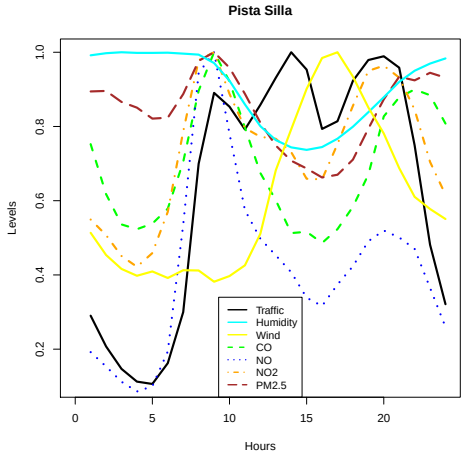

Figure 6. Average daily traffic intensity and pollution parameters measured in Pista Silla station.

trations can cause acute and chronic effects to anyone, especially those with sensitivity.

The ranges of risk shown by the application from the predicted values of the four pollutants are based on the recommendations of the Directive 2008/50/EC (European Comission, 2008). The variable as NOx (oxides of nitrogen) refers to NO or NO2, since the normative establishes the same limits for both levels.

- **Green level:** [NOx] $<$ 14.0 μg/m3 $\wedge$ [CO] $<$ 30.0 mg/m3 $\wedge$ [PM 2.5] $<$ 7.5 μg/m3.
- **Yellow level:** We establish medium risk (yellow level) if the levels do not satisfy the conditions of the green level and the red level.
- **Red level:** [NOx] $\geq$190.0 μg/m3 $\vee$ [CO] $\geq$ 55.0 mg/m3 $\vee$ [PM 2.5] $\geq$ 25.0 μg/m3

4.3. Risk warnings

Airvlc mobile application can be configured to send warnings to users if the device is near to a zone (200 meters approximately) where a high risk level is predicted. These warnings can be personalised by the user in different ways.

For example, the user can establish personal limits for warnings or modify the range of distance for the detection of high risk levels of pollutant concentration. Obviously, the user needs to allow the application to know the actual GPS location of the device

In the case of the web application, given that here it is more complex to know the exact location of the user, we adopt a different strategy. We are working in an automated warning system where the user needs to fix a set of areas, and then the system sends an electronic email whenever a dangerous situation (high risk level by default) is detected.

5. Related work

A wide number of works employs machine learning techniques or statistical approaches for predicting pollution levels. A classical work is (Yi & Prybutok, 1996). In this paper, the authors propose ozone prediction models. Specifically, they develop a neural network model for forecasting daily maximum ozone levels and compare it to previous approaches by regression, and Box-Jenkins ARIMA. The results show that the neural network model improves the performance of the regression and Box-Jenkins ARIMA models tested. Neural networks models have been widely

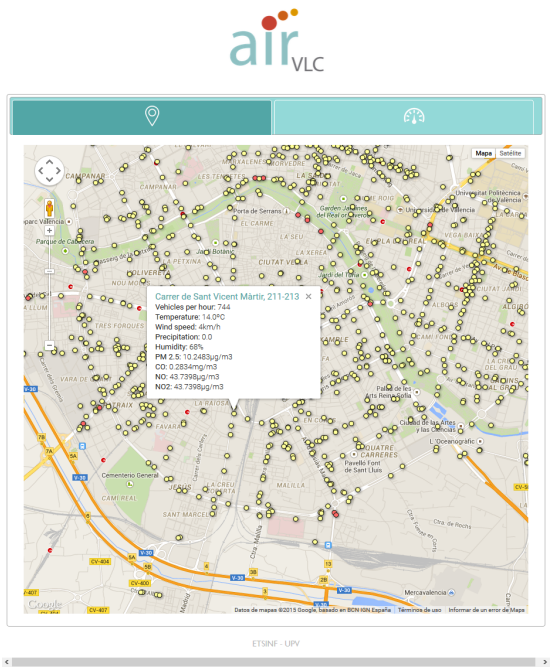

Figure 7. Airvlc application. Real-time pollution map.

Figure 8. Frame where the user can introduce specific locations to know the predicted levels of pollution.

employed in this field, a review of these approaches can be found in (Khare & Nagendra, 2006).

A more related work is (Karppinen et al., 2000a). Here the authors propose a modelling system for predicting the traffic volumes, emissions from stationary and vehicular sources, and atmospheric dispersion of pollution in an urban area. They employ four monitoring stations in the Helsinki metropolitan area in 1993. The paper compares the predicted NOx and NO2 concentrations with the results of an urban air quality monitoring network. The agreement of model predictions was better for the two suburban monitoring stations, compared with two urban stations. Some applications of these models are introduced in (Karppinen et al., 2000b). A similar work for the city of Izmir in Turkey is (Elbir, 2003). Here, the authors compare The CALMET meteorological model and its puff dispersion model CALPUFF for predicting dispersion of the sulphur dioxide emissions from industrial and domestic sources.

Another related work, and in this case very recent, is (Donnelly et al., 2015). This paper presents a model for real time air quality forecasts. The predictions are concentrated in nitrogen dioxide (NO2) and they are used to estimate air quality 48 hours in advance. The model is based on a multiple linear regression which uses linearised factors describing variations in concentrations together with meteorological parameters and persistence as predictors.

Our comparison of regression techniques obtains similar conclusions to the work presented in (Singh et al., 2013). In this study, principal components analysis (PCA) is performed to identify air pollution sources. From the extracted features, tree based ensemble learning models are induced to predict the urban air quality of Lucknow (India) together with the air quality and meteorological databases for a period of five years.

6. Conclusions

Air pollution can decrease life expectancy since contamination rises the risk of suffering respiratory diseases. Although policies motivating the reduction of emissions of pollutant particles have been introduced in the last years, many cities frequently still present risky levels of air pollution. In these situations, the reduction of the exposure to ambient air pollution is highly recommended. In this work, we have presented Airvlc, an application that predicts in real-time the levels of four dangerous pollutants in a wide set of points in the city of Valencia. The system is able to predict these pollution levels by applying regression models trained from data containing information traffic intensity, persistence of pollutants and meteorological parameters. Airvlc can be a useful tool for avoiding risky locations in terms of air pollution.

As future work we propose the integration of the application in middleware platforms such as Fi-Ware[5], this could help to extend the applicability of the system to other cities or regions. We also are interested in the incorporation of additional features in order to improve the prediction models: wind direction, sand storms, forest wildfires and agricultural burnings... Finally, the use of the tool for the recommendation of routes that minimise the exposure to air pollution.

[5] http://www.fiware.org/

Acknowledgments

We thank the anonymous reviewers for their comments, which have helped to improve this paper significantly. We are also grateful to Ajuntament de València, InnDEA València and specially to Ramón Ferri, Ruth López and Paula Llobet for their help in providing traffic data. This work was supported by the REFRAME project, granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net (CHIST-ERA), and funded by the Ministerio de Economía y Competitividad in Spain (PCIN-2013-037). It also has been partially supported by the EU (FEDER) and the Spanish MINECO project ref. TIN2013-45732-C4-01 (DAMAS), and by Generalitat Valenciana ref. PROMETEOII/2015/013 (SmartLogic).

References

Donnelly, Aoife, Misstear, Bruce, and Broderick, Brian. Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmospheric Environment*, 103:53–65, 2015.

Elbir, Tolga. Comparison of model predictions with the data of an urban air quality monitoring network in izmir, turkey. *Atmospheric Environment*, 37(15):2149–2157, 2003.

European Comission. Directive 2008/50/ec of the european parliament on ambient air quality and cleaner air for europe. http://ec.europa.eu/environment/air/quality/legislation/directive.htm, 2008.

Hornik, Kurt, Buchta, Christian, and Zeileis, Achim. Open-source machine learning: R meets Weka. *Computational Statistics*, 24(2):225–232, 2009. doi: 10.1007/s00180-008-0119-7.

Karppinen, A, Kukkonen, J, Elolähde, T, Konttinen, M, and Koskentalo, T. A modelling system for predicting urban air pollution:: comparison of model predictions with the data of an urban measurement network in helsinki. *Atmospheric Environment*, 34(22):3735–3743, 2000a.

Karppinen, A, Kukkonen, J, Elolähde, T, Konttinen, M, Koskentalo, T, and Rantakrans, E. A modelling system for predicting urban air pollution: model description and applications in the helsinki metropolitan area. *Atmospheric Environment*, 34(22):3723–3733, 2000b.

Khare, Mukesh and Nagendra, SM Shiva. *Artificial neural networks in vehicular pollution modelling*, volume 41. Springer, 2006.

Koenker, Roger. *quantreg: Quantile Regression*, 2015. URL http://CRAN.R-project.org/package=quantreg. R package version 5.11.

Liaw, Andy and Wiener, Matthew. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL http://CRAN.R-project.org/doc/Rnews/.

Meyer, David, Dimitriadou, Evgenia, Hornik, Kurt, Weingessel, Andreas, and Leisch, Friedrich. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. URL http://CRAN.R-project.org/package=e1071. R package version 1.6-4.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

Singh, Kunwar P, Gupta, Shikha, and Rai, Premanjali. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment*, 80:426–437, 2013.

The European Environment Agency . Soer 2015 — the european environment — state and outlook 2015. http://www.eea.europa.eu/soer, 2015.

US Environmental Protection Agency . Particulate matter (pm) regulations. http://www.epa.gov/airquality/particlepollution/index.html, 2015.

Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

Wilker, Elissa H., Preis, Sarah R., Beiser, Alexa S., Wolf, Philip A., Au, Rhoda, Kloog, Itai, Li, Wenyuan, Schwartz, Joel, Koutrakis, Petros, DeCarli, Charles, Seshadri, Sudha, and Mittleman, Murray A. Long-Term Exposure to Fine Particulate Matter, Residential Proximity to Major Roads and Measures of Brain Structure. *Stroke*, April 2015. doi: 10.1161/strokeaha.114.008348. URL http://dx.doi.org/10.1161/strokeaha.114.008348.

World Health Organisation. Public health, environmental and social determinants of health. http://www.who.int/phe/health_topics/outdoorair/databases/health_impacts/en/, 2015.

Yi, Junsub and Prybutok, Victor R. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution*, 92(3):349–357, 1996.