



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

 etsinf

Escola Tècnica
Superior d'Enginyeria
Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica

Universitat Politècnica de València

Evaluación sobre la calidad del aire en España

Trabajo Final de Máster

Máster en Big Data Analytics

Autor: Antonio Avia Antúnez
Tutor: María José Ramírez Quintana
2017 - 2018

Resumen

En el presente trabajo se presenta un análisis sobre la contaminación del aire en España, centrado principalmente en metales y gases de efecto invernadero. Este estudio pretende demostrar como de importante es la reducción de la emisión de esos gases sobre todo y principalmente debido a la actividad humana. A partir de aquí se presentan modelos predictivos que pueden determinar como serán esas emisiones de cara a un futuro. Se mostrarán datos a nivel de municipio, provincia y comunidad autónoma.

Palabras clave: contaminación, emisiones, salud, actividad humana.

Resum

Al present treball es presenta un análisis sobre la contaminació de l'aire a Espanya, centrat principalment en metals i gasos d'efecte hivernacle. Aquest estudi pretén demostrar com d'important és la reducció de l'emissió d'eixos gasos sobretot i principalment a causa de l'activitat humana. A partir d'ací es presenten models predictius que poden determinar com seràn eixes emissions de cara a un futur. Es mostraran dades a nivell de municipi, província i comunitat autònoma.

Paraules clau: contaminació, emissions, salut, activitat humana.

Resum

This thesis presents an analysis on air pollution in Spain, mainly focused on metals and greenhouse gases. This study aims to demonstrate how important it is to reduce the emission of these gases especially and mainly due to human activity. From here, predictive models are presented that can determine how these emissions will be in the future. Data will be displayed at the level of municipality, province and autonomous community.

Key words: pollution, emissions, health, human activity.



Tabla de contenidos

1	Introducción	5
1.1.	Objetivos.....	7
1.2.	Estructura del documento.....	7
2	Conceptos previos	8
2.1.	Introducción a la Minería de Datos.....	8
2.2.	Tipos de Tareas de Minería de Datos	12
	Tareas Predictivas	13
	Tareas Descriptivas	13
2.3.	Técnicas de Aprendizaje Supervisado	14
	Regresión Lineal	14
	Árbol de Decisión	14
	Random Forest	15
	Red Neuronal	16
2.4.	Técnicas de Aprendizaje No Supervisado	17
	Correlaciones	18
	Agrupamiento.....	18
	Gaussian Mixture Models	19
3	Comprensión de los datos	20
3.1.	Estructura de los Datos	21
3.2.	Análisis Descriptivo de datos Horarios, Diarios e Irregulares.....	24
	Datos Horarios	24
	Datos Diarios.....	25
	Datos Irregulares	28
4	Análisis Descriptivo de los datos unificados	29
4.1.	Pre-procesos	29
4.2.	Contaminación según altitud de la estación sobre el nivel del mar	30
4.3.	Contaminación por municipios de costa.....	31
4.4.	Contaminación por tamaño de municipios.....	32
4.5.	Contaminación por tipo de área.....	33
4.6.	Contaminación por fuente de emisión	34
4.7.	Gráficos de niveles de contaminación por municipios	35
4.8.	Estudio sustancias más comunes	39



5	Análisis basado en agrupamiento usando K-Means	39
5.1.	Estudios a realizar	40
5.2.	Construcción del conjunto de datos	40
5.3.	Técnicas para calcular el valor de K óptimo.....	43
	Método Elbow.....	43
	Análisis de Silueta.....	43
	Agrupamiento por Provincia, Municipio, Contaminante y Trimestre	44
	Ejecución de K-Means por Provincia, Municipio, Contaminante y Trimestre	49
	Agrupamiento por Provincia, Municipio, Gases y Trimestre	58
	Ejecución de K-Means por Provincia, Municipio, Gases y Trimestre.....	61
	Agrupamiento por Provincia, Municipio, Metales y Trimestre	66
	Ejecución de K-Means por Provincia, Municipio, Metales y Trimestre.....	68
6	Análisis Predictivo	74
6.1.	Elección Municipio para Análisis Predictivo	74
	Definición de Vista Minable	77
	Técnicas de Predicción aplicadas.....	81
7	CONCLUSIÓN.....	85
7.1.	Comentarios	¡Error! Marcador no definido.
8	Bibliografía	88
9	ANEXO I. Descripción de los contaminantes	89
10	ANEXO II. Evaluación de la Calidad el Aire	93
11	ANEXO III. Gráficos de contaminantes por ubicación geográfica	97
12	ANEXO IV. Trabajos relacionados	103



1 Introducción

En tiempos actuales, sin duda alguna existe una gran preocupación sobre el cambio climático, de cómo se ha originado, cuáles son sus causas, cómo va a ser su evolución y también, en gran medida, qué consecuencias tiene a medio y largo plazo.

Se habla del ascenso de la temperatura media mundial respecto a datos históricos registrados.

Se estima que el nivel del mar pueda subir más de un metro a finales de este siglo, lo que haría desaparecer poblaciones enteras y provocaría grandes migraciones de personas, que además quedarían sin recursos. También se habla del aumento de la acidez del mar, afectando a grandes poblaciones de peces y de arrecifes de coral, rompiendo el ecosistema del arrecife haciéndolo desaparecer.

Tanto el aumento de los períodos de sequía, como el aumento de las lluvias torrenciales incontrolables, el desbordamiento de los ríos, la contaminación de las aguas y la lluvia ácida amenazan los cultivos y terrenos de pastos de todo el mundo, contribuyendo a la desaparición de especies animales y vegetales, viéndose afectado por lo tanto también el ser humano.

Otros fenómenos naturales no menos importantes a tener en cuenta son el aumento de la frecuencia de volúmenes en erupción, incendios forestales, tsunamis, terremotos, tornados, huracanes y cualquier otro fenómeno natural que contribuya a la destrucción de zonas naturales, agrícolas y habitadas.

Esta crisis climática afectará a nuestra salud, aumentando el riesgo de alergias, enfermedades, virus, epidemias y pandemias.

Aunque existen análisis mucho más amplios y profundos, algunos de ellos incluidos en el Anexo IV, la mayoría de los científicos y expertos coinciden en que las causas principales del cambio climático y el calentamiento global son:

Transporte contaminante

El 40% de las emisiones provienen de vehículos a motor, sean terrestres, embarcaciones acuáticas o aéreas.

Edificios que requieren una rehabilitación energética

El 36% de los gases emitidos proviene de edificios y viviendas con carencias o puntos a mejorar en relación a:

- *Aislamiento.* Un correcto aislamiento térmico reduce al 50% el consumo de energía
- *Estanqueidad.* Evitar fugas de aire y sellar huecos entre ventanas y paredes puede reducir entre un 30% y un 50% el consumo de energía



- *Ventilación.* Una ventilación eficiente reduce en un 90% la demanda de frío o de calor por aparatos eléctricos

Generación de residuos

Cada persona genera aproximadamente un kilo y medio de basura al día.

El 60% de esta basura se trata de envases, embalajes, bolsas de plástico y derivados.

La gestión de estos residuos se había centrado principalmente en enviarlos a vertederos y plantas incineradoras, siendo esta solución no sostenible por diferentes motivos:

- Son un riesgo para el medioambiente, los seres vivos y la salud de las personas por las emisiones que estos procesos producen
- No reduce el consumo de materias primas y energía
- No actúa sobre el modelo de consumo para cambiarlo y/o mejorarlo

Agricultura y ganadería. Sistema alimentario no sostenible

Los sistemas actuales de producción de alimentos son *ineficientes e insostenibles* y son responsables del 60% de la pérdida de biodiversidad a nivel global y del 24% de las emisiones de gases de efecto invernadero.

La deforestación para agricultura, la sobreexplotación de caladeros y la contaminación de suelos y acuíferos son algunas de las causas directas de la pérdida de biodiversidad, a las que hay que sumar el uso de combustibles fósiles, teniendo, por tanto, un impacto directo en la seguridad alimentaria.

Derroche de energía

Existen gran cantidad de medidas que podemos tomar en nuestro día a día para reducir un gasto innecesario de energía en nuestro hogar, lugares de trabajo, zonas de ocio, etc.

Deforestación

La deforestación tiene un impacto tanto en su entorno más cercano como en el resto del planeta. Los árboles transforman el CO₂ en oxígeno, siendo el CO₂ uno de los gases principales que emitimos. Al eliminar bosques la concentración de este gas aumenta en la atmósfera.

Reciclar, reutilizar, consumir de forma responsable, generar menos residuos, ..., en definitiva, llevar a cabo pequeñas acciones por parte de cada uno puede provocar un gran cambio a nivel global para revertir esta situación.

1.1. Objetivos

El presente documento forma parte del *Trabajo Final de Máster del Máster en Big Data Analytics por la Universidad Politécnica de Valencia* y pretende hacer un pequeño y modesto acercamiento de cómo de importante es centrar todos los esfuerzos y recursos para hacer visible cómo la actividad humana está contribuyendo a la emisión de gases de efecto invernadero y que, por lo tanto, no hacen más que acelerar un calentamiento global que en un futuro cada vez menos incierto se está cumpliendo.

Todo el estudio realizado se basa en mediciones proporcionadas como datos públicos por el Ministerio para la Transición Ecológica.

Estas mediciones forman parte de las comunicaciones que España realiza anualmente a la Comisión Europea sobre las mediciones realizadas en diferentes municipios de las comunidades autónomas españolas siguiendo determinadas directivas y normativas en materia de calidad del aire, centrándose en gases y metales considerados de efecto invernadero (véase el Anexo II).

Existe una red de estaciones de medición por todo el país, donde cada Comunidad Autónoma gestiona y recoge los datos, incluidas Baleares y Canarias.

Con el objetivo de extraer el máximo de información de los datos ofrecidos por el ministerio, los siguientes capítulos del documento se dividen en dos grandes grupos:

- Estudio descriptivo de las mediciones obtenidas del Ministerio para la Transición Ecológica

Dentro del estudio descriptivo, a su vez, existen tres enfoques para este tipo de estudio:

- Estudio descriptivo de datos horarios, diarios e irregulares
- Estudio descriptivo unificando esos datos horarios, diarios e irregulares en un único conjunto de datos
- Estudio descriptivo utilizando técnicas de agrupación o clasificación. K-Means
- Por último, un estudio predictivo sobre una sustancia contaminante en concreto para un municipio determinado

1.2. Estructura del documento

Esta memoria está estructurada como sigue:

- En el capítulo 2 se explica en qué consisten las técnicas de minería de datos, los diferentes tipos de tareas que existen y las técnicas de análisis descriptivo y predictivo que se pueden aplicar.
- El capítulo 3 detalla la estructura de datos con la que se trabaja y se hace un estudio específico de cada fuente de datos: diarios, horarios e irregulares.

- En el capítulo 4 se unifica la información en un único conjunto de datos y se hace un estudio descriptivo de diferentes tipologías.
- El capítulo 5 se centra en un análisis descriptivo con *K-Means*.
- En el capítulo 6 se detalla el análisis predictivo realizado.
- Los capítulos finales contienen un apartado de conclusiones y anexos que aportan información adicional a este trabajo:
 - Descripción de los contaminantes estudiados: fuentes de emisión, aplicaciones en la actividad humana y efectos para la salud
 - Medidas de evaluación de la calidad del aire de la Comisión Europea
 - Gráficos adicionales de los estudios realizados
 - Referencias de otros trabajos relacionados

2 Conceptos previos

2.1. Introducción a la Minería de Datos

Para poder estudiar en detalle los datos de los que disponemos vamos a desarrollar lo que se conoce como un estudio de *Minería de Datos*. Un proyecto estándar de *Minería de Datos* proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software.

La metodología de *Minería de Datos* contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de modelos. Este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no sólo no acaba una vez se halla el modelo idóneo (ya que después se requiere un despliegue y un mantenimiento), sino que está relacionado con otros proyectos, y es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él.

El ciclo de vida de un proyecto de *Minería de Datos*, de acuerdo a la metodología *CRISP-DM* (*Wirth & Hipp, 2000*) consta de seis fases:

Fase I. Comprensión del Negocio

Esta fase inicial se enfoca en la comprensión de los objetivos del proyecto. Después se convierte este conocimiento de los datos en la definición de un problema de *Minería de Datos* y en un plan preliminar diseñado para alcanzar los objetivos. Una descripción de cada una de las principales tareas que componen esta fase es la siguiente:

- Determinar los objetivos del negocio. Tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar *Minería de Datos* y definir los criterios de



éxito. En cuanto a los criterios de éxito, estos pueden ser de tipo cualitativo, en cuyo caso un experto en el área de dominio, califica el resultado del proceso, o de tipo cuantitativo.

- Evaluación de la situación. En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de *Minería de Datos*, considerando aspectos tales como: ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de *Minería de Datos* ?, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de *Minería de Datos*.
- Determinación de los objetivos del proceso de *Minería de Datos*. Tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de *Minería de Datos*, desarrollando un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada paso.

Fase II. Estudio y Comprensión de los datos

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta:

- Recolección de datos iniciales. Destinada a la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo, elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.
- Descripción de los datos. Después de adquiridos los datos iniciales, estos deben ser descritos. Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.
- Exploración de datos. A continuación, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. La salida de esta tarea es un informe de exploración de los datos.
- Verificación de la calidad de los datos. En esta tarea, se efectúan verificaciones sobre los datos, para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea de este punto es asegurar la completitud y corrección de los datos.

Fase III. Análisis de los datos y selección de características

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan:

- Estructuración de los datos. Incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes.
- Integración de los datos. Involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.
- Formateo de los datos. Realización de transformaciones sintácticas de los datos sin modificar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de *Minería de Datos* en particular, como por ejemplo la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.).

Fase IV. Modelado

En esta fase se seleccionan las técnicas de modelado más apropiadas para el proyecto de *Minería de Datos* específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada al problema
- Disponer de datos adecuados
- Cumplir los requisitos del problema
- Tiempo adecuado para obtener un modelo
- Conocimiento de la técnica

Previamente al modelado de los datos, se debe determinar un método de evaluación de los modelos que permita establecer el grado de bondad de ellos. Después de concluir estas tareas genéricas, se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo. Una descripción de las principales tareas de esta fase es la siguiente:

- Selección de la técnica de modelado. Esta tarea consiste en la selección de la técnica de *Minería de Datos* más apropiada al tipo de problema a resolver. Se debe considerar el objetivo principal del proyecto y la relación con las herramientas de *Minería de Datos* existentes



- Generación del plan de prueba. Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo. Por ejemplo, en una tarea supervisada de *Minería de Datos* como la clasificación, es común usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.
- Construcción del Modelo. Después de seleccionada la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.
- Evaluación del modelo. En esta tarea, los ingenieros de *Minería de Datos* interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en *Minería de Datos* aplican sus propios criterios (seguridad del conjunto de prueba, perdida o ganancia de tablas, etc...).

Fase V. Evaluación. Obtención de resultados

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse, además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error. Se pueden emplear múltiples herramientas para la interpretación de los resultados. Las matrices de confusión son muy empleadas en problemas de clasificación y consisten en una tabla que indica cuantas clasificaciones se han hecho para cada tipo, la diagonal de la tabla representa las clasificaciones correctas. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo. Las tareas involucradas en esta fase del proceso son las siguientes:

- Evaluación de los resultados. En los pasos de evaluación anteriores, se trataron factores tales como la exactitud y generalidad del modelo generado. Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual, el modelo sea deficiente, o si es aconsejable probar el modelo, en un problema real si el tiempo y restricciones lo permiten. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional.
- Proceso de revisión. Se refiere a calificar al proceso entero de *Minería de Datos*, con objeto de identificar elementos que pudieran ser mejorados.
- Determinación de futuras fases. Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría pasarse a la fase siguiente, en caso contrario podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros



parámetros. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de Minería de Datos.

Fase VI. Implementación

En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso. Generalmente un proyecto de *Minería de Datos* no concluye en la implantación del modelo, pues se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento. Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados. Las tareas que se ejecutan en esta fase son las siguientes:

- Plan de implementación. Para implementar el resultado de *Minería de Datos* en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior implementación.
- Monitorización y Mantenimiento. Si los modelos resultantes del proceso de *Minería de Datos* son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.
- Informe Final. Es la conclusión del proyecto de *Minería de Datos* realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto. Revisión del proyecto: En este punto se evalúa qué fue lo correcto y qué lo incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar.

2.2. Tipos de Tareas de Minería de Datos

Un tipo de tarea de *Minería de Datos* es un tipo de problema que se desea resolver usando un conjunto de datos recogidos para tal fin y usando un algoritmo del aprendizaje automático o de la estadística.

Existen dos tipos de tareas: predictivas y descriptivas. Entre las tareas predictivas encontramos la clasificación y la regresión, mientras que el agrupamiento (*clustering*), es la tarea descriptiva más extendida.



Tareas Predictivas

Será un conjunto de datos en el que cada ejemplo o instancia es una tupla de valores. En las tareas predictivas el objetivo es predecir el valor de uno de los atributos, el cual recibe el nombre de variable objetivo o “*clase*”. El resto de los atributos se denominan variables de entrada, a partir de las cuales trataremos de determinar el valor de la variable objetivo.

En las tareas predictivas se emplean diversas técnicas estadísticas de modelización y aprendizaje automático para reunir toda la información y elaborar predicciones de cara al futuro.

Normalmente, se utilizan datos históricos para crear un modelo matemático que captura las tendencias importantes y relaciones, tanto en el conjunto de datos estructurados como no estructurados. Este modelo predictivo se usa entonces con los datos actuales para predecir lo que pasará a continuación, o bien para sugerir acciones a llevar a cabo con el fin de obtener resultados óptimos.

Existen gran número de métodos predictivos basados en diferentes algoritmos matemáticos, que se adaptarán mejor dependiendo de con qué objetivo se analicen los datos.

Dependiendo del tipo de la variable a predecir se distinguen dos tipos de tareas predictivas:

Clasificación

Se caracteriza porque la variable a predecir es nominal o categórica, es decir, puede tomar valor de entre un conjunto finito de valores.

Regresión

Se caracteriza porque la variable a predecir es numérica.

Tareas Descriptivas

Las tareas descriptivas proporcionan información sobre las relaciones entre los datos y sus características, permiten extraer patrones, tendencias y regularidades para *DESCRIBIR* y comprender mejor los datos.

Ayuda a identificar *asociaciones* y *dependencias* entre diferentes atributos cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente es relativamente alta.

Aplicable a datos con características de tipo enumerado, valores limitados o fijos. Los objetivos de este tipo de modelos son:

- Descubrir grupos de muestras similares, agrupaciones naturales entre los datos, también conocido como *clustering*
- Determinar cómo se distribuyen los datos en el espacio de entrada mediante la estimación de densidades de probabilidad



- Reducción de la dimensionalidad. Proyectar los datos desde un espacio de alta dimensionalidad a dos o tres dimensiones para visualizar los datos

Por ejemplo, se investigará si existe relación entre el tamaño de las ciudades y la cantidad de algunos contaminantes.

2.3. Técnicas de Aprendizaje Supervisado

Para resolver las tareas predictivas se aplican técnicas del aprendizaje supervisado, el cual engloba aquellos algoritmos que buscan aprender una función f que permita mapear una instancia $x = (x_1, \dots, x_n)$ a una determinada clase c

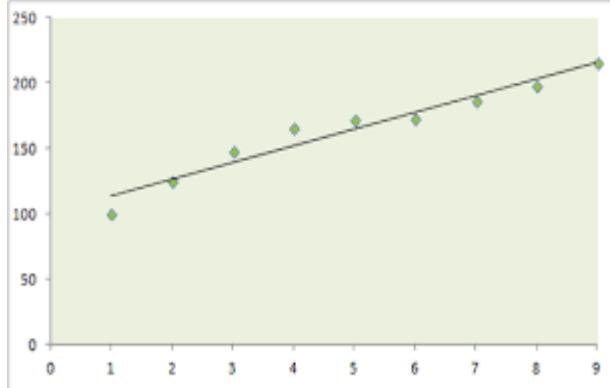
$$f(x_1, \dots, x_n) \rightarrow c$$

Para ello, requiere disponer de un conjunto de datos etiquetados, es decir, un conjunto de datos $D = \{(x_1, \dots, x_n, c)\}$, del que se conoce el valor de la variable “clase”. En general, el conjunto de datos originales se parte en dos: un conjunto de entrenamiento para generar los modelos, y un conjunto de test para evaluar los modelos.

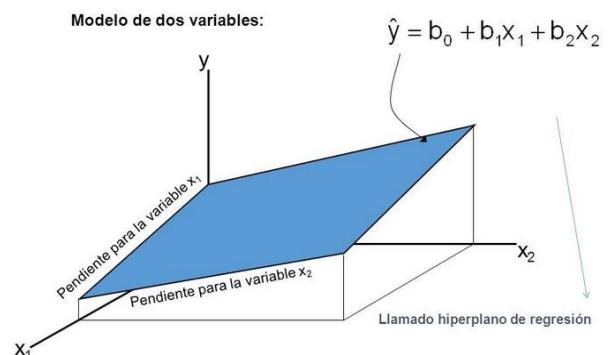
Regresión Lineal

El análisis de regresión lineal se utiliza para explicar una determinada variable, digamos Y, en función de una variable X, o bien en función de una combinación lineal de varias variables X_1, X_2, \dots, X_k . En el primer caso se trata de una Regresión Univariante mientras que en el segundo sería una Regresión Multivariante:

Regresión Lineal Univariante



Regresión Lineal Multivariante (de 2 variables)

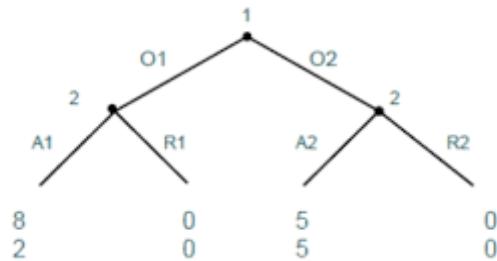


Árbol de Decisión

Se trata de un conjunto de condiciones organizadas en una estructura jerárquica tipo árbol. Está formado por nodos de decisión que realizan un test sobre el valor de una variable X_i .



Las diferentes respuestas que se pueden dar generan otros nodos de decisión. El árbol se va construyendo, de forma que se crea un modelo que *pronostica valores de la variable de interés* basada en valores de otras variables. Se trata de un *modelo de aprendizaje supervisado*.



Random Forest

Se trata de una técnica basada en la combinación de multitud de *árboles de decisión*. La idea es generar distintos árboles usando conjuntos de datos diferentes y después combinar las decisiones de los árboles de base. Para ello, se sigue el procedimiento conocido como *bagging*, que consiste en generar subconjuntos de entrenamiento seleccionando aleatoriamente y con reemplazamiento una muestra de ejemplos del conjunto de datos original del mismo tamaño que éste.

Adicionalmente, para evitar que haya un clasificador muy influyente junto con otros moderadamente influyentes, *Random Forest* hace una selección aleatoria de las variables de entrada para ser consideradas en cada nodo interno de cada árbol. De esta forma, se consigue *decorrelacionar* los árboles, por lo que su agregación consigue una mayor reducción de la varianza.

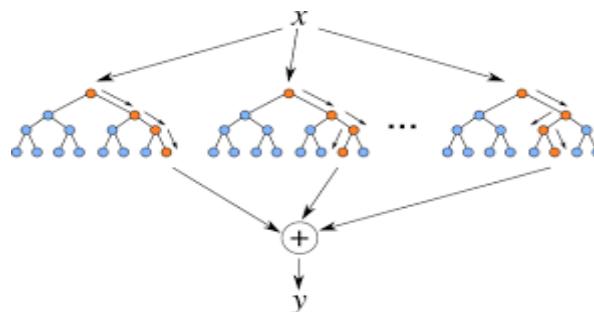
Otra de las ventajas de *Random Forest* es que no sufre problemas de *overfitting* al aumentar el número de árboles creados en el proceso. Alcanzado un determinado número, la reducción de *test error* se estabiliza.

Se trata de un modelo de *aprendizaje supervisado* y es útil para regresión y clasificación, sirviendo también como técnica para reducción de la dimensionalidad.

Al generar múltiples árboles, cada árbol da una clasificación, votando por una clase.

El resultado es la clase con mayor número de votos en todo el bosque (forest).

Para regresión, se toma el promedio de las salidas (predicciones) de todos los árboles.



Red Neuronal

Una Red Neuronal es un modelo matemático inspirado en el comportamiento biológico de las neuronas y en cómo se organizan formando la estructura del cerebro. Las redes neuronales intentan aprender, mediante ensayos repetidos, cómo organizarse mejor a sí mismas para conseguir maximizar la predicción.

Un modelo de red neuronal se compone de nodos, que actúan como input, output o procesadores intermedios. Cada nodo se conecta con el siguiente conjunto de nodos mediante una serie de trayectorias ponderadas. Basado en un paradigma de aprendizaje, el modelo toma el primer caso, y toma inicial basada en las ponderaciones. Se evalúa el error de predicción y modifica las ponderaciones para mejorar la predicción, a continuación, se evalúa el siguiente caso con las nuevas ponderaciones y se modifican para mejorar la predicción de los casos ya evaluados, el ciclo se repite para cada caso en lo que se denomina la fase de preparación o evaluación. Cuando se ha calibrado el modelo, con la muestra test se evalúan los resultados globales.

El elemento básico de una red neuronal es un nodo. Es la unidad de procesamiento que actúa en paralelo con otros nodos de la red. Es similar a las neuronas del cerebro humano: acepta input y genera output. Los nodos aceptan input de otros nodos. La primera tarea del nodo es procesar los datos de entrada creando un valor resumen que es la suma de todas las entradas multiplicadas por sus ponderaciones. Este valor resumen se procesa a continuación mediante una función de activación para generar una salida que se envía al siguiente nodo del sistema.

Se considera una red neuronal la ordenación secuencial de tres tipos básicos de nodos o capas: nodos de entrada, nodos de salida y nodos intermedios (capa oculta o escondida).

Los nodos de entrada se encargan de recibir los valores iniciales de los datos de cada caso para transmitirlos a la red.

Los nodos de salida reciben entradas y calculan el valor de salida (no van a otro nodo). En casi todas las redes existe una tercera capa denominada oculta.

Este conjunto de nodos utilizados por la red neuronal, junto con la función de activación, posibilita a las redes neuronales representar fácilmente las relaciones no lineales, que son muy problemáticas para las técnicas multivariantes.

Un ejemplo de modelo neuronal con n entradas consta de:

- Un conjunto de entradas x_1, \dots, x_n (*Input Layer*)
- Un conjunto de pesos w_1, \dots, w_n correspondientes a cada entrada
- Un conjunto de nodos correspondientes a una capa oculta (*Hidden Layer*)
- Una función de agregación \sum
- Una función de activación f
- Una salida $Y (z_k)$

Las entradas son el estímulo que la neurona artificial recibe del entorno que la rodea, y la salida es la respuesta a tal estímulo. La neurona puede adaptarse al medio circundante y aprender de él

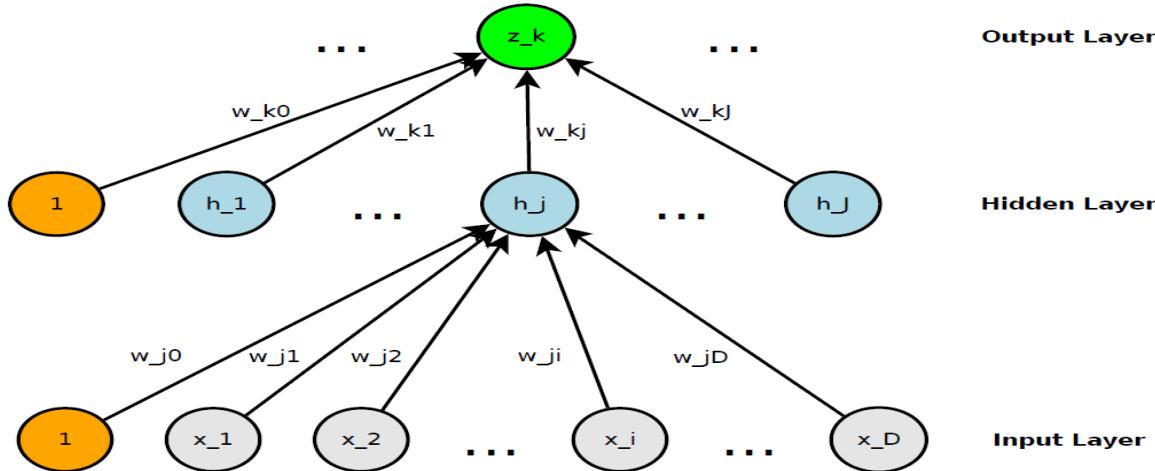


modificando el valor de sus pesos sinápticos, y por ello son conocidos como los parámetros libres del modelo, ya que pueden ser modificados y adaptados para realizar una tarea determinada.

En este modelo, la salida neuronal Y está dada por:

$Y = f(\sum n_i = I w_i x_i)$, siendo la función de activación elegida de acuerdo a la tarea realizada por la neurona.

Su representación gráfica sería:



Diferentes técnicas descriptivas y predictivas mencionadas se han utilizado en este estudio y serán detalladas en los siguientes capítulos.

En el caso de las técnicas predictivas, se predecirán los valores de contaminación de una de las sustancias en función de valores históricos y otros parámetros adicionales.

2.4. Técnicas de Aprendizaje No Supervisado

En el aprendizaje no supervisado, la finalidad es modelar la estructura o distribución subyacente en los datos a partir de datos no etiquetados, es decir, donde ningún atributo se ha designado como objetivo, de manera que se lleva a cabo una exploración de los datos a partir de las variables de entrada. La finalidad de este tipo de aprendizaje es descubrir las diferentes categorías que describen las características de los datos no etiquetados.

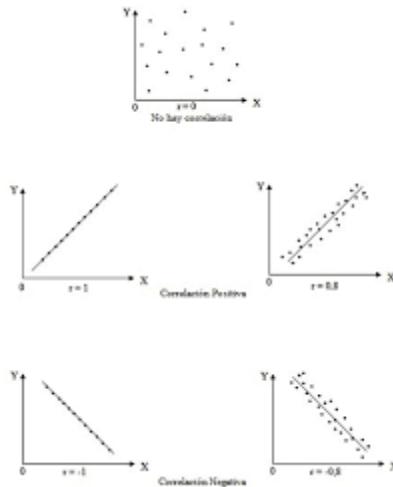
Las técnicas de aprendizaje no supervisado trabajan comúnmente con la siguiente estructura de datos:



	X_1	...	X_i	...	X_n
$\mathbf{x}^{(1)}$	$x_1^{(1)}$...	$x_i^{(1)}$...	$x_n^{(1)}$
\dots	\dots	\dots	\dots	\dots	\dots
$\mathbf{x}^{(j)}$	$x_1^{(j)}$...	$x_i^{(j)}$...	$x_n^{(j)}$
\dots	\dots	\dots	\dots	\dots	\dots
$\mathbf{x}^{(N)}$	$x_1^{(N)}$...	$x_i^{(N)}$...	$x_n^{(N)}$

Correlaciones

Permite conocer la existencia de datos relacionados entre sí.



Aplicable a datos numéricos. Por ejemplo, según nuestro caso de estudio, observaremos el comportamiento y las relaciones entre las diferentes moléculas de *Nitrógeno*.

Agrupamiento

También conocido como *clustering*, merece un apartado específico puesto que será una técnica que se utilizará en este estudio, más concretamente el algoritmo conocido como *K-Means*.

Su objetivo es identificar grupos de individuos que son similares entre sí de acuerdo a una cierta noción de similitud.

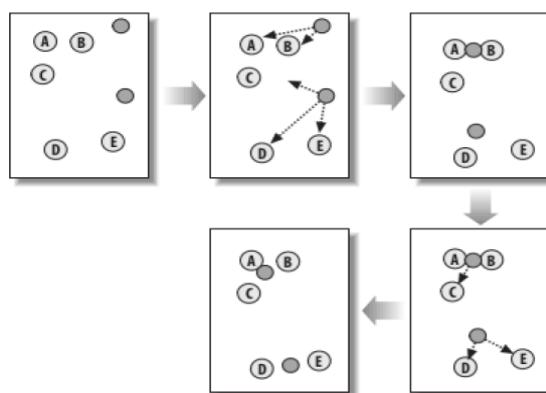
Por definición, *K-Means* es un *algoritmo no supervisado*, ya que no sabemos cuántos grupos existen inicialmente, y su funcionamiento se basa en la identificación, dentro de un conjunto de datos, de grupos o *clústers* en un espacio d-dimensional, es decir, el objetivo es encontrar K prototipos que sean representativos de cada grupo, de forma que cada muestra será asignada al *clúster* a cuyo prototipo esté más próxima.



Aunque pueden utilizarse otras métricas, el más común es el cálculo de la media basado en la distancia euclídea entre cada par de puntos de ese espacio d-dimensional, y será ésta la métrica utilizada en este estudio.

La ejecución del algoritmo consta de tres pasos:

- **Elección de un valor de K óptimo.** Para determinar el valor de K óptimo existen diferentes técnicas que nos ayudarán a tomar esta decisión, esos métodos son el *Método Elbow* o del codo y el *Método de Análisis de Silueta*. En siguientes los capítulos se detallarán.
- **Inicialización:** Una vez escogido el número de grupos K , se establecen K centroides en el espacio de los datos. En general, no existe un modo exacto de determinar el valor K inicial, pero se puede estimar con aceptable precisión siguiendo diferentes técnicas que se explicarán a continuación.
- **Asignación de objetos a los centroides:** Los objetos o elementos serán los puntos de cualquier color que no sea el amarillo y todos los del mismo color corresponderán a elementos del mismo grupo, tendrán características similares. En cada iteración del algoritmo, las muestras se asignarán a su centroide más cercano.
- **Actualización de centroides:** Los centroides serán los puntos amarillos y en cada iteración del algoritmo se irán ajustando (moviendo) respecto a la media de las muestras más cercanas.



Gaussian Mixture Models

Adicionalmente, junto con agrupamiento K-Means, se utilizará también esta técnica.

Los modelos de mezcla gaussiana (GMM) suponen que hay un cierto número de distribuciones gaussianas, y cada una de estas distribuciones representa un grupo. Por lo tanto, un modelo de mezcla gaussiana tiende a agrupar los puntos de datos que pertenecen a una única distribución, a un único grupo.

Supongamos que tenemos tres distribuciones gaussianas: GD1, GD2 y GD3. Éstas tienen un cierto valor medio (μ_1, μ_2, μ_3) y varianza ($\sigma_1, \sigma_2, \sigma_3$) respectivamente. Para un conjunto de puntos de datos, nuestro GMM identificaría la probabilidad de que cada punto de datos pertenezca a cada una de estas distribuciones, por lo tanto, son modelos probabilísticos y utilizan el enfoque de agrupación suave para distribuir los puntos en diferentes agrupaciones o distribuciones gaussianas.

Los modelos de *distribuciones gaussianas* se construyen en base a dos parámetros o criterios específicos del algoritmo:

- AIC: Akaike Information Criterion
- BIC: Bayesian Information Criterion

En algunos casos, se utilizan en las primeras fases de un estudio para identificar relaciones en los datos que puedan ayudar a descubrir patrones en los datos que se repiten más porque forman *clústers*, es decir, se concentran o presentan mayor densidad en ciertas regiones del espacio multidimensional.

En nuestro caso de estudio éste será el objetivo, servirá como herramienta de apoyo para determinar un número óptimo de *clústers K*, que será el valor que minimiza el *AIC* o el *BIC*.

3 Comprensión de los datos

Los datos públicos proporcionados por el Ministerio para la Transición Ecológica según la dirección <https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/evaluacion-datos/datos/> contienen mediciones de las sustancias contaminantes de determinados gases y metales considerados de efecto invernadero. Esas sustancias son las siguientes:

Sustancia	Símbolo	Unidad de medida
Dióxido de Azufre	SO ₂	µg/m ³
Monóxido de Carbono	CO	mg/m ³
Monóxido de Nitrógeno	NO	µg/m ³
Dióxido de Nitrógeno	NO ₂	µg/m ³
Partículas en suspensión < 2.5 µM	PM _{2.5}	µg/m ³
Partículas en suspensión < 10 µM	PM ₁₀	µg/m ³
Óxidos de Nitrógeno	NO _x	µg/m ³
Ozono	O ₃	µg/m ³
Arsénico	As	ng/m ³
Plomo (PM10)	Pb	µg/m ³



Benzoapireno (PM10)	BAP	ng/m ³
Cadmio (PM10)	Cd	ng/m ³
Benceno	C ₆ H ₆	μg/m ³
Níquel (PM10)	Ni	ng/m ³

Indicar, que no todas las provincias tienen el mismo número de estaciones e incluso que la calidad de la información registrada por las estaciones tampoco es la misma, encontrándose en algunos casos poca cantidad de datos o incluso gran cantidad de *valores nulos*, aspecto que hace más complicado un análisis. El Anexo I incluye una descripción más detallada de estos contaminantes.

La tipología de cada estación es diferente, obteniendo mediciones en zonas urbanas, suburbanas y rurales. En estas mediciones también se especifica el origen de la fuente de emisión que lo origina, que puede ser relacionado con el tráfico de los medios de transporte, de zonas industriales cercanas o de zonas donde no predomina ninguna de las anteriores.

Para poder comprender de qué información disponemos, se hace necesario utilizar determinadas técnicas de análisis adecuadas para la extracción del conocimiento de esos datos, cómo se comportan, como se relacionan entre sí, si siguen un determinado patrón en función de otros parámetros, etc.

Esos patrones o tendencias se utilizarán, tanto para describir los datos, como para predecir comportamientos futuros de los mismos.

3.1. Estructura de los Datos

Los datos proporcionados se encuentran clasificados en 3 grandes grupos y cada fichero corresponde a las mediciones registradas de una única sustancia:

Datos Horarios

Fichero	Sustancia medida	Símbolo Químico
C6H6_HH_YYYY.csv	Benceno	C ₆ H ₆
CO_HH_YYYY.csv	Monóxido de Carbono	CO
NO2_HH_YYYY.csv	Dióxido de Nitrógeno	NO ₂
NOx_HH_YYYY.csv	Óxidos de Nitrógeno	NOx
NO_HH_YYYY.csv	Monóxido de Nitrógeno	NO
O3_HH_YYYY.csv	Ozono	O ₃
PM10_HH_YYYY.csv	Partículas en Suspensión 10 diámetro	PM ₁₀
PM25_HH_YYYY.csv	Partículas en Suspensión 2.5 diámetro	PM _{2.5}
SO2_HH_YYYY.csv	Dióxido de Azufre	SO ₂



Datos Diarios

Fichero	Sustancia medida	Símbolo Químico
As_DD_yyyy.csv	Arsénico	As
B(a)P_DD_yyyy.csv	Benzoapireno	BAP
Cd_DD_yyyy.csv	Cadmio	Cd
Ni_DD_yyyy.csv	Níquel	Ni
Pb_DD_yyyy.csv	Plomo	Pb
PM10_DD_yyyy.csv	Partículas en Suspensión 10 diámetro	PM ₁₀
PM2.5_DD_yyyy.csv	Partículas en Suspensión 2.5 diámetro	PM _{2.5}

Datos Irregulares

Fichero	Sustancia medida	Símbolo Químico
C6H6_yyyy_Irreg.csv	Benceno	C ₆ H ₆

Contiene únicamente mediciones de la sustancia *Benceno* realizadas en intervalos de tiempo que pueden abarcar valores dentro de un mismo mes o en meses diferentes. En este caso solo hay datos del contaminante para Castilla La Mancha y Andalucía.

A diferencia de los datos horarios y diarios, la periodicidad del muestreo de estos datos es variable por abarcar diferentes intervalos de tiempo, de ahí su denominación como *irregulares*.

yyyy = *Todos los ficheros horarios, diarios e irregulares corresponden a datos de los años 2016 y 2017. En próximos capítulos se ampliará esta información.*

Los tres ficheros presentan la siguiente estructura en común:

Código/Identificador asociado a una estación	Compuesto por	Código Provincia (código INE)	
		Código Municipio (código INE)	
		Identificador interno de la estación	
Magnitud		Código del contaminante o sustancia medida	
Punto de muestreo		Representa una serie de datos y se utiliza para, dentro de una misma ubicación, realizar una medición con diferentes equipos	

De forma específica según el tipo de fichero a tratar tenemos:

Datos horarios

Año	Mes	Dia	Magnitud	H0	H1	...	H23

Donde H0 ... H23 corresponde a las mediciones registradas desde las 0 hasta las 23h de ese día de la sustancia correspondiente.



Datos diarios

Año	Mes	Magnitud	D1	D2	D3	...	D31

Donde D1 ... D31 corresponde a las mediciones registradas desde el día 1 hasta el día 31 del mes indicado de la sustancia correspondiente. Las mediciones de los meses con menos días no presentarán valores.

Datos Irregulares

Fecha Inicio	Fecha Fin	Magnitud

Fecha inicio y fin en formato DD/MM/YYYY de la medición dentro del mismo año para la sustancia correspondiente. Puede abarcar varios días dentro de un mismo mes o incluso en meses diferentes.

También se dispone de información general de las estaciones y su clasificación.

Información General

Nombre	Fecha Inicio	Fecha Fin	Red	CCAA	Provincia	Municipio	Latitud / Longitud	Altitud

- Nombre de la estación
- Fecha de Inicio de la actividad de la estación
- Fecha de Fin de la actividad de la estación
- Red a la que pertenece. Normalmente se indica la comunidad autónoma a la que pertenece
- Comunidad Autónoma donde está ubicada la estación
- Provincia donde está ubicada la estación
- Municipio donde está ubicada la estación
- Latitud / Longitud en coordenadas geográficas en grados
- Altitud de la estación sobre el nivel del mar en metros

Clasificación y Tipología

- Según el tipo de área donde están ubicadas se clasifican en:

Urbanas

Estaciones ubicadas en zonas edificadas



Suburbanas

Estaciones ubicadas en zonas con presencia de zonas edificadas separadas por zonas no urbanizadas (lagos, bosques, terreno agrícola, ...)

Rurales

Estaciones ubicadas en zonas que no cumplen los criterios anteriores

- Según el tipo principal de la fuente de emisión que la influye se clasifican:

De Tráfico

El nivel de contaminación viene determinado por emisiones de vehículos de calles o carreteras próximas

Industriales

El nivel de contaminación viene determinado por emisiones de una zona industrial

De Fondo

El nivel de contaminación no está determinado por ninguna fuente de emisión predominante

A continuación, en el siguiente capítulo se mostrará una amplia descripción de los datos, tomando como referencia diferentes atributos y variables incluso estudios desde diferentes puntos de vista

3.2. Análisis Descriptivo de datos Horarios, Diarios e Irregulares

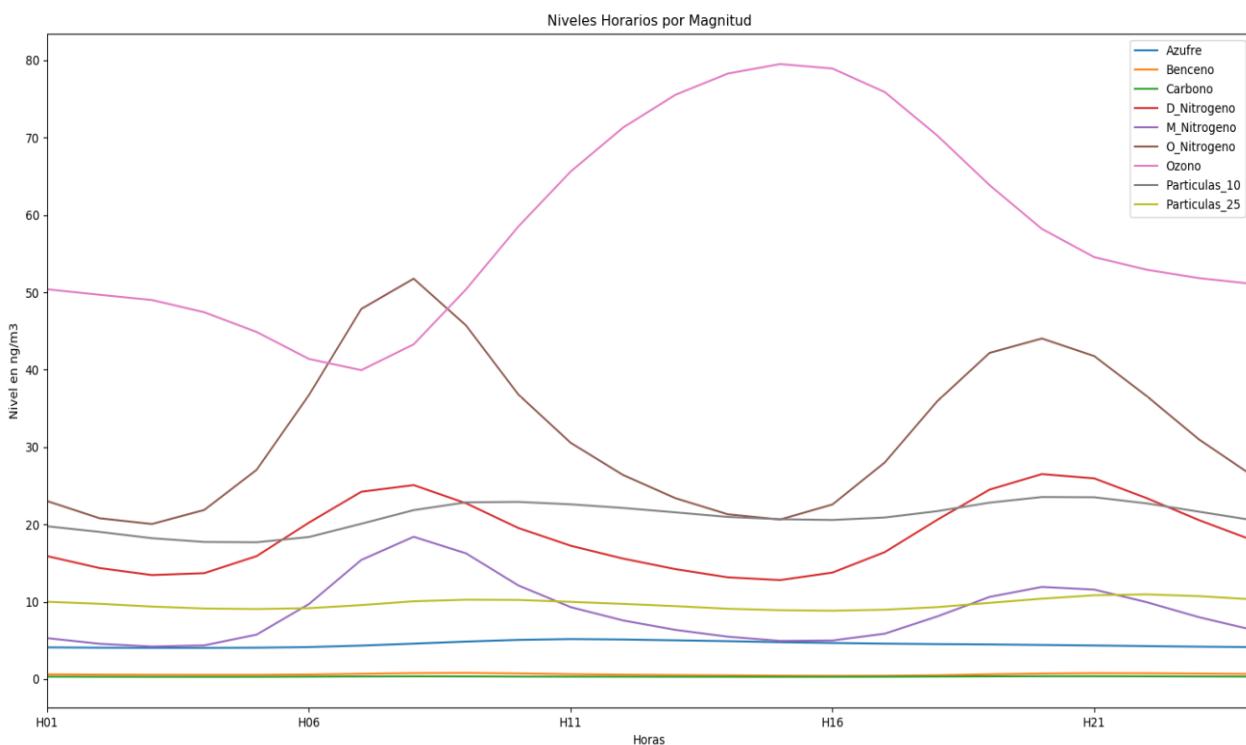
Para el análisis descriptivo se han utilizado datos registrados de 2017.

Datos Horarios

Corresponde a los valores medios de las 24 mediciones realizadas por horas para un mismo día de todas las estaciones para las diferentes sustancias.

Tras unificar todos los datos horarios, se muestran los resultados de todo el año agrupados por horas, según se muestra en los siguientes gráficos:





Lo que más destaca es que el *Ozono* es la sustancia predominante, sobre todo en las horas de más sol, por lo tanto, a mayor radiación solar, mayor presencia de *Ozono*.

Después del *Ozono*, la sustancia que más está presente es el *Nitrógeno* en sus diferentes moléculas (*Dióxido, Monóxido y Óxido*). Los niveles más elevados se presentan entre las 7 y las 9h de la mañana por un lado y las 19 y las 21h de la tarde. Esta circunstancia está relacionada con el mayor desplazamiento de vehículos para acudir/salir de los centros de trabajo, puesta en marcha/parada de maquinaria o de procesos en zonas industriales. Los altos niveles de *Ozono* se correlacionan con bajos niveles de *Nitrógeno* y viceversa.

La presencia de partículas en suspensión en términos generales suele ser elevada, presentando mayores valores las *Partículas en Suspensión < 10µM*.

También destaca la presencia de forma constante del *Dióxido de Azufre* a lo largo de un día con valores promedios de 4 µg/m³.

El *Benceno* y el *Carbono* presentan valores prácticamente ínfimos.

Por último, para las sustancias como el *Níquel, Cadmio, Arsénico, Plomo y Benzoapireno* no se dispone de información horaria registrada por las estaciones.

Datos Diarios

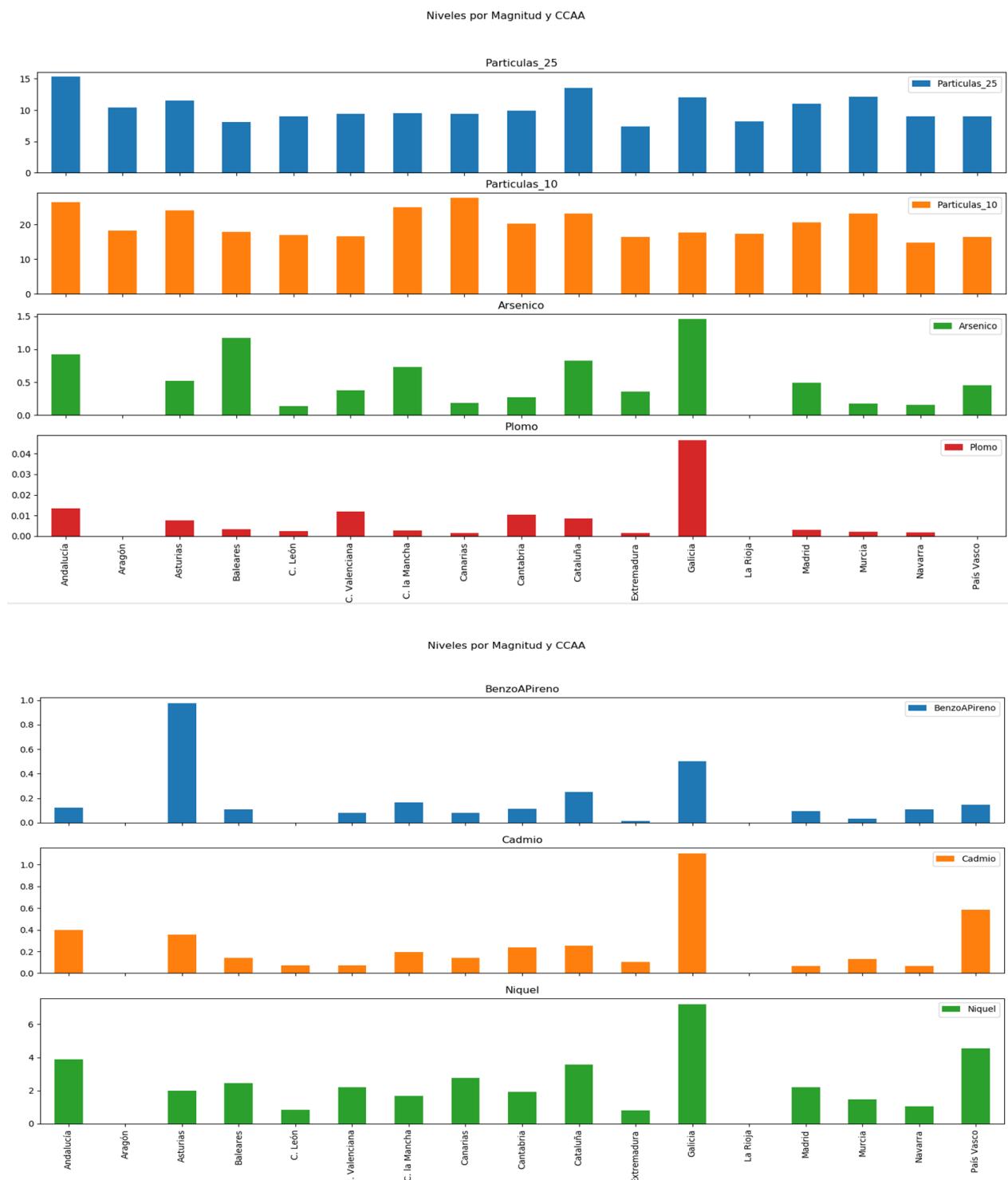
Corresponde a los valores medios de las 31 mediciones realizadas para todos los meses de todas las estaciones para las diferentes sustancias. Los meses que tienen menos días no presentan medición en esos días concretos.



Tras unificar todos los datos diarios, se muestran los resultados de dos formas:

Agrupación Anual

Se muestran los valores medios de los dos años por Comunidad Autónoma.



Las siguientes sustancias: *Azufre*, *Carbono*, *Nitrógeno* (en sus diferentes moléculas) y *Ozono* no presentan valores diarios registrados.

Según los resultados, la CCAA con mayor contaminación es Galicia, además con grandes diferencias respecto al resto.

Aragón y La Rioja, sólo presentan mediciones de *Partículas en Suspensión*.

La presencia de *Partículas en Suspensión* es un rasgo general para todas las CCAA.

Con independencia de Galicia, existe presencia importante de *Arsénico* en Andalucía, Cataluña e Islas Baleares.

Los niveles de Plomo, en general, son bajos, aunque en Andalucía y Comunidad Valenciana superan el 0.01 $\mu\text{g}/\text{m}^3$.

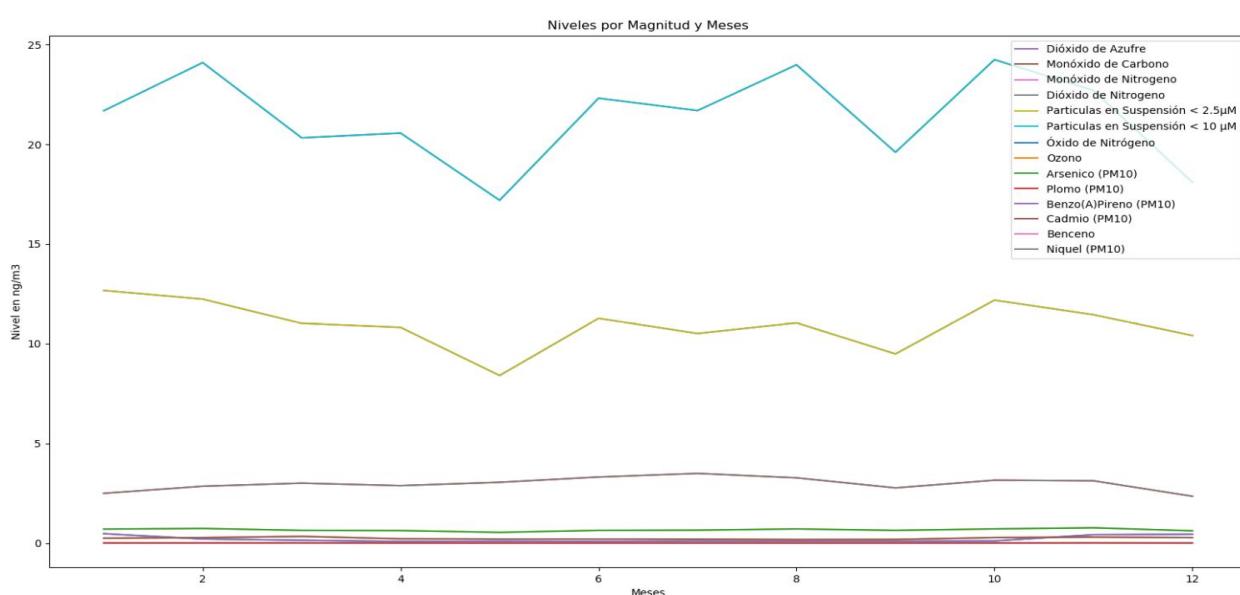
Los niveles de *Benzoapireno* son muy altos en Asturias, 1 ng/m^3 , si lo comparamos con el resto.

Existe una presencia importante de *Cadmio* en el País Vasco seguido de Andalucía.

Por último, los niveles de Níquel en Andalucía, Cataluña y País Vasco también son elevados.

Agrupación Mensual

Se muestran los valores medios mensuales de los dos años.



Según los resultados, queda claro que, a lo largo de un año, la sustancia que mayor presencia tiene son las *Partículas en Suspensión*, con una gran diferencia respecto a las demás. De las *Partículas en Suspensión < 10 μM* , los meses de mayor concentración son febrero, agosto y octubre. Los de menor mayo y diciembre.



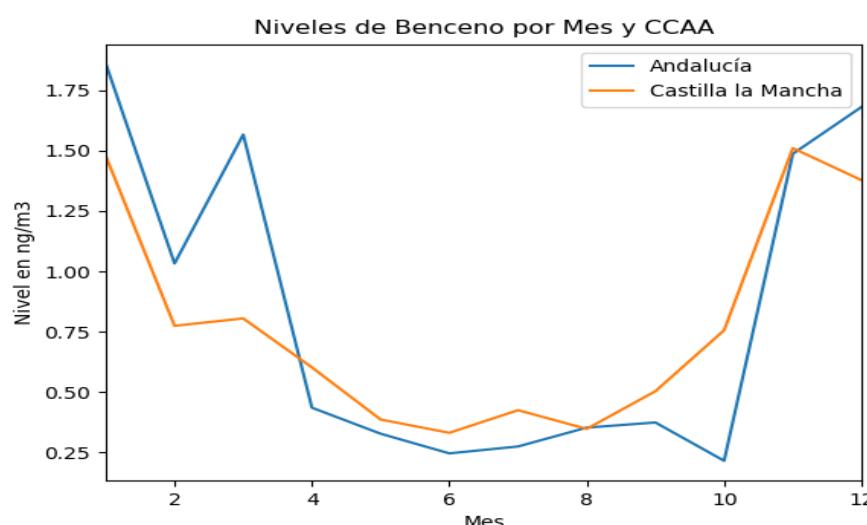
De las *Partículas en Suspensión < 2.5 μM*, los meses de mayor concentración son enero y octubre. Los de menor son mayo y septiembre.

Detrás de las *Partículas en Suspensión* destaca el *Dióxido de Nitrógeno*, aunque su valor es constante a lo largo del año.

El resto de las sustancias presenta niveles constantes muy bajos, aunque de este grupo, el que mayor presencia muestra es el *Arsénico*.

Datos Irregulares

El resultado del estudio muestra que son valores únicamente de la sustancia *Benceno* para las CCAA de Andalucía y Castilla la Mancha.



Como se observa, el comportamiento del *Benceno* en las dos CCAA es similar.

Lo más destacable en Andalucía es la oscilación entre los meses de febrero, marzo y abril, por un lado, y de octubre, noviembre y diciembre por otro.

En Castilla la Mancha la oscilación no es tan brusca, aunque también queda marcada de enero a febrero, por una parte, y de agosto a noviembre por otra.

El *Benceno* está compuesto químicamente por moléculas de *Carbono e Hidrógeno*, y como tal, sus valores irán directamente relacionados con la presencia de estas sustancias.

Adicionalmente, las emisiones de tráfico y sistemas de calefacción son otra fuente importante de emisión de este contaminante, por lo tanto, parece claro que el comportamiento de esta sustancia sea de tipo estacional, ya que, en primavera y verano, la actividad de tráfico y de sistemas de calefacción es inferior, así como que las altas temperaturas ayudan a volatilizarlo.



4 Análisis Descriptivo de los datos unificados

A diferencia del punto anterior, en este capítulo se pretende unificar toda la información de datos diarios, horarios e irregulares en un único conjunto de datos para hacer un análisis desde otro punto de vista, como es la localización geográfica de las estaciones de medición, altura sobre el nivel del mar, cercanía a la costa, tamaño del municipio, tipo de área y fuente de emisión predominante de las sustancias contaminantes.

Los datos analizados corresponden al 2017.

Para ello se han preparado los datos aplicando diferentes pre-procesos.

4.1. Pre-procesos

De forma individual, para los **datos diarios, horarios e irregulares**, se genera la media anual para una provincia, municipio y sustancia contaminante dada.

Aunque el análisis descriptivo se ha centrado en el año 2017, más adelante se explica en qué casos se utilizan también los datos del 2016. En este sentido, se tiene la siguiente información relativa a valores de contaminantes recibidos como Nulos:

Datos Diarios	2016	2017
Total de Filas del conjunto de datos	6645	10744
Elementos Nulos	239	526
% de Nulos sobre el conjunto total	3.59 %	4.89 %

El motivo por el que los valores Nulos de los *datos horarios* no están indicados es porque el estudio descriptivo no se centra en un estudio de franjas horarias, por lo tanto, para trabajar con esta información el primer paso que se llevó a cabo fue transformar esos datos en una media diaria.

Respecto a los *datos irregulares*, por definición, únicamente tienen información de períodos concretos porque la periodicidad de sus muestreos es variable, por lo tanto, no podemos considerar que presenten valores Nulos.

Unificación de datos

Tras unificar los tres conjuntos de datos anteriores (diarios, horarios e irregulares) en un único conjunto de datos, se calcula de nuevo la media de esa sustancia contaminante por provincia y municipio para todo el año.

En este momento ya tenemos un conjunto de datos único:



Datos Diarios	Datos Horarios	Datos Irregulares
Provincia	Provincia	Provincia
Municipio	Municipio	Municipio
Sustancia	Sustancia	Sustancia
Media anual para todos los días	Media anual para todas las horas	Media anual para todos los intervalos de fechas



Conjunto Único
Provincia
Municipio
Sustancia
Media [media anual datos diarios, media anual datos horarios, media anual datos irregulares]

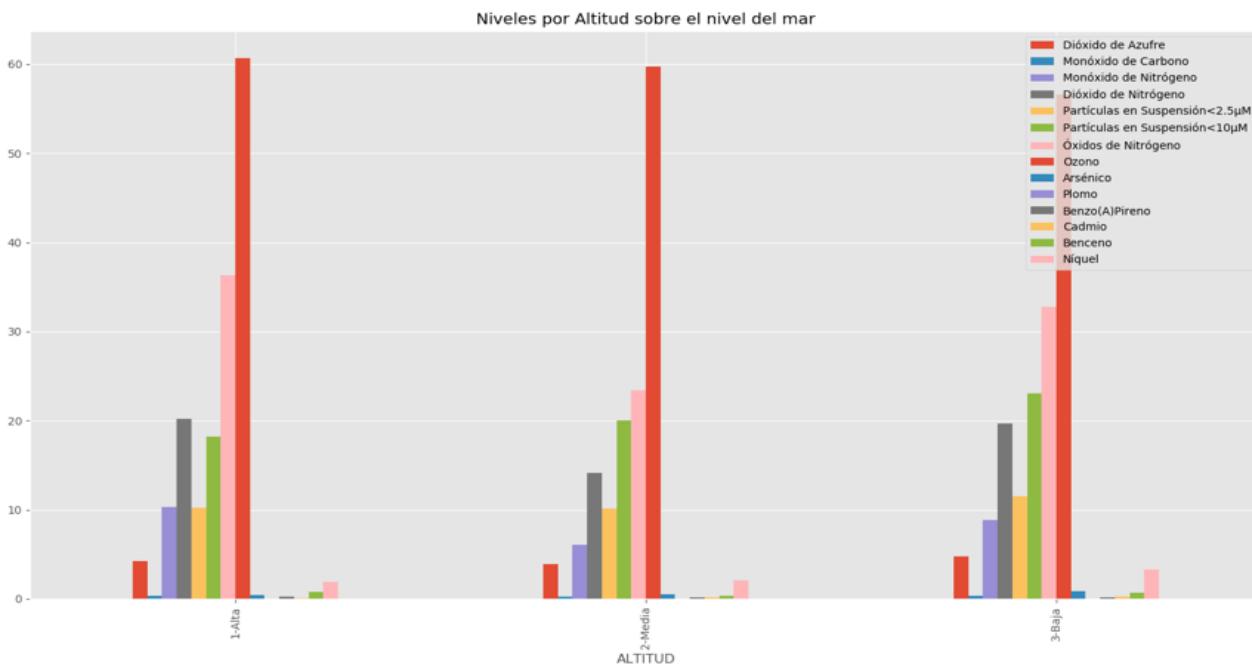
A partir de esta estructura unificada ya podemos añadir otras variables de estudio para observar cómo se comportan los datos de contaminación con respecto a otros factores externos, permitiendo así conocer mejor su comportamiento.

4.2. Contaminación según altitud de la estación sobre el nivel del mar

El objetivo de este estudio es determinar si es importante la altura sobre el nivel del mar respecto a la contaminación existente, estableciendo una comparativa entre municipios altos, medios y bajos, para ello se ha obtenido para todos los municipios presentes en el conjunto de datos, sus alturas respectivas sobre el nivel del mar.

Se asignan las siguientes categorías según altura sobre el nivel del mar del municipio donde se ubica la estación que registra las mediciones:

- *Alta*. Para estaciones con altura sobre el nivel del mar igual o superior a 600 m
- *Media*. Para estaciones con altura sobre el nivel del mar igual o superior a 200 m y menor a 600 m
- *Baja*. Para estaciones con altura sobre el nivel del mar inferior a 200 m



En primera instancia, parece que no existen grandes diferencias según la altura de la estación.

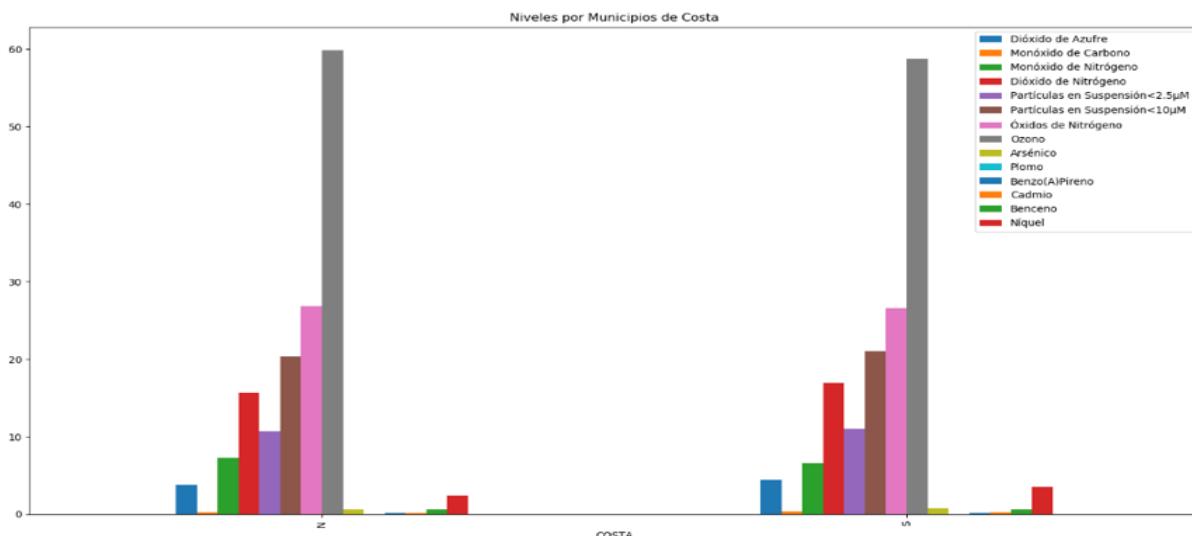
- Los valores del *Ozono* son lógicos, a mayor altura, mayores niveles de concentración.
- Respecto al *Nitrógeno*, se concentran mayores niveles en las capas altas y bajas que en las intermedias
- Sin embargo, sí se observa que las *Partículas en Suspensión* se concentran más en las capas bajas y que por lo tanto a mayor altura, menor particulado contaminante.
- Respecto a la presencia de metales contaminantes, se ve ligeramente que existe mayor concentración cuanto más baja la altura sobre el nivel del mar es.

4.3. Contaminación por municipios de costa

El objetivo de este estudio es determinar si el hecho de que un municipio sea de costa o no influye en la concentración de sustancias contaminantes, estableciendo una comparativa entre municipios de costa y de interior, para ello se ha obtenido para todos los municipios sus distancias a la costa.

Se asignan las siguientes categorías según distancia del municipio al mar donde se ubica la estación:

- Se considera municipio de costa si está a una distancia menor o igual a 5 km del mar
- No se considera municipio de costa si está a una distancia superior a 5 km del mar



Como se puede ver existen diferencias muy pequeñas, apenas perceptibles, entre municipios de costa y municipios de interior.

Los valores de *Dióxido de Azufre*, *Dióxido de Nitrógeno*, *Partículas en Suspensión*, *Arsénico* y *Níquel* son ligeramente superiores en municipios de costa.

Aun así, no se puede determinar que el atributo municipio de costa sea decisivo e influya directamente en la concentración de sustancias contaminantes, ya que en términos generales los valores son muy similares.

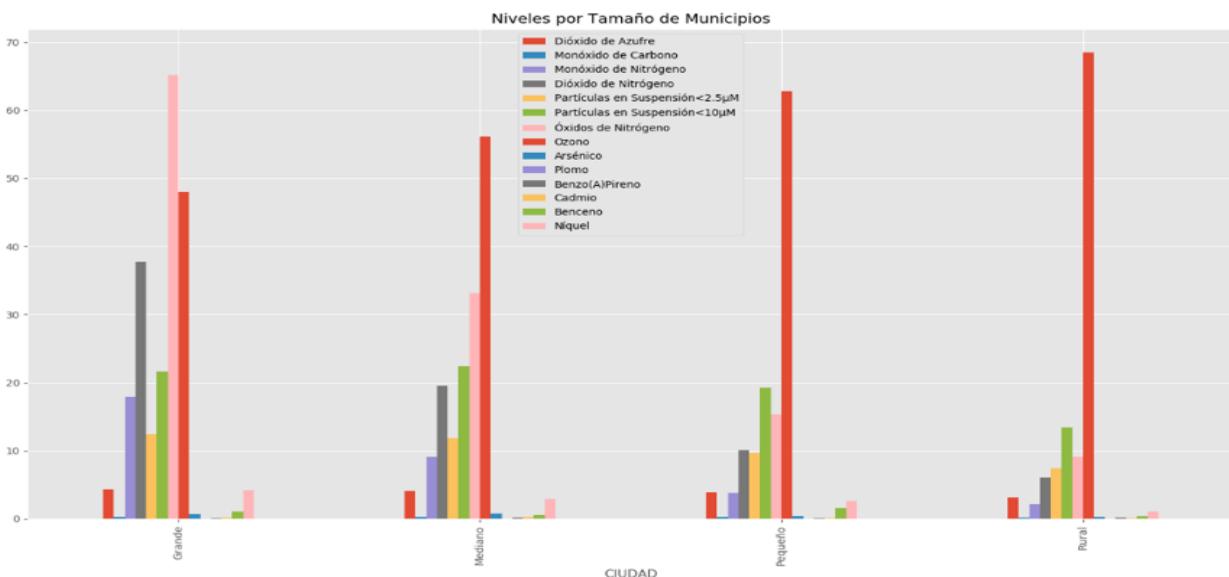
4.4. Contaminación por tamaño de municipios

El objetivo de este estudio es determinar si la población de un municipio influye en la concentración de sustancias, estableciendo 4 categorías de tamaño de población, para ello se ha obtenido para todos los municipios presentes en el conjunto de datos, sus poblaciones respectivas.

En este sentido se puede presuponer que desde el punto de vista de zonas de tráfico de vehículos y de zonas industriales tengan gran influencia los municipios de mayor población,

Se asignan las siguientes categorías según el municipio donde está ubicada la estación:

- *Grande*. Se trata de municipios con una población mayor o igual a 700.000 habitantes
- *Mediano*. Se trata de municipios con una población entre 10.000 y 700.000 habitantes
- *Pequeño*. Se trata de municipios con una población entre 2.500 y 10.000 habitantes
- *Rural*. Se trata de municipios con una población inferior a 2.500 habitantes



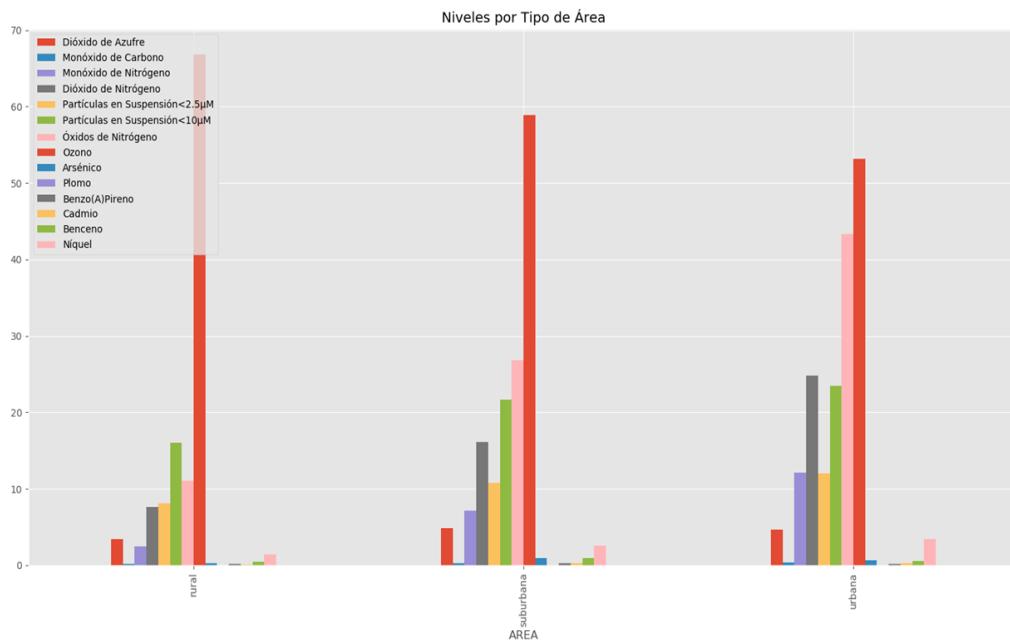
A mayor tamaño de población, mayor concentración de sustancias contaminantes, con la excepción del *Ozono*, que presenta valores más altos en zonas rurales y más despobladas.

En este sentido sí que se puede concluir que el tamaño de población de los municipios es importante e influyente en los valores de los contaminantes.

4.5. Contaminación por tipo de área

El objetivo de este estudio es determinar si el tipo de área donde está ubicada la estación influye en la concentración de sustancias, estableciendo 3 categorías. Esta información se ha obtenido a través del conjunto de datos de la clasificación de estaciones indicado en el *Capítulo 3.1*.

- *Urbana*. Se trata de zonas edificadas de forma permanente
- *Suburbana*. Se trata de zonas edificadas separadas por zonas no urbanizadas, tales como bosques, lagos, tierras agrícolas, etc
- *Rural*. Son aquellas zonas que no cumplen ninguna de las otras dos categorías



La tendencia de los contaminantes es bastante clara, presentando niveles ascendentes según el orden de tipo de área rural, suburbana y urbana.

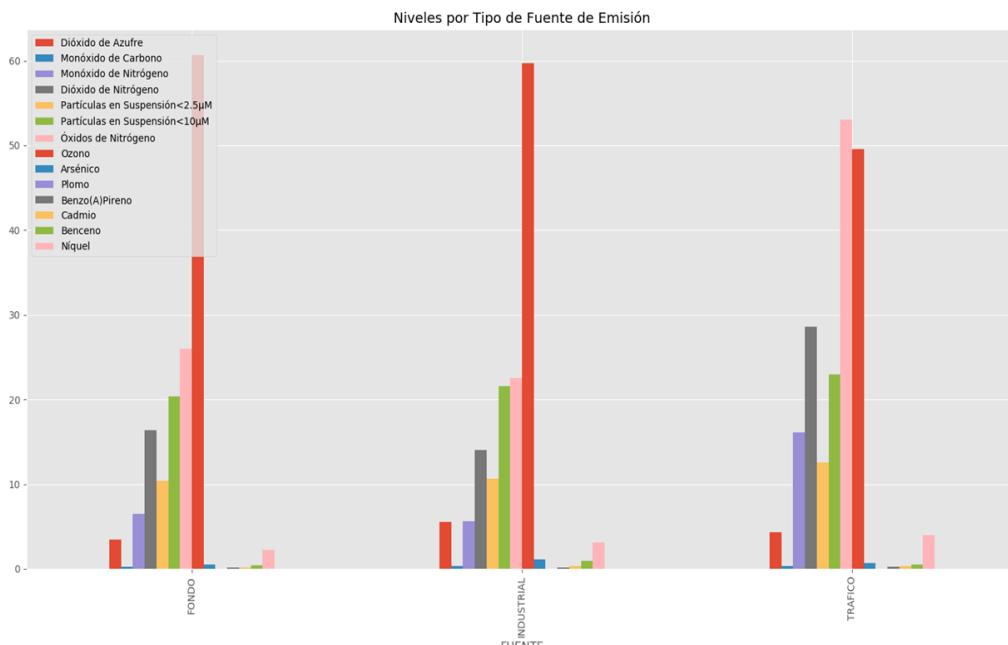
Los valores del *Nitrógeno* (en sus diferentes moléculas) presentan de forma muy presente niveles muy altos para fuentes de emisión de tipo tráfico.

Por otro lado, el efecto del *Ozono* es menor en zonas urbanas que en zonas suburbanas, se entiende que debido al efecto de *sustentación* que las ciudades ejercen respecto a zonas más alejadas de los núcleos urbanos.

4.6. Contaminación por fuente de emisión

El objetivo de este estudio es determinar si la fuente de emisión de la sustancia es importante, estableciendo 3 categorías. Esta información se ha obtenido a través del conjunto de datos de la tipología de estaciones indicado en el *Capítulo 3.1*.

- *De tráfico*. El nivel de contaminación está determinado por emisiones de vehículos.
- *Industrial*. El nivel de contaminación está determinado por emisiones de zonas industriales.
- *De fondo*. El nivel de contaminación no está determinado por zonas industriales ni de tráfico, no existen fuentes de emisión predominantes.



Como se observa, destaca principalmente el *Ozono*, siendo incluso superior la fuente de emisión industrial a la fuente de tráfico.

Evidentemente en fuentes de emisión de tipo tráfico los valores de las diferentes moléculas de *Nitrógeno* son los más elevados de todos, así como que se ve una línea ascendente en los valores del *Níquel* y de *material particulado*.

Si se comparan los gráficos por tipo de área y por fuente de emisión se ven ciertas similitudes entre zonas rurales y emisiones de fondo, zonas suburbanas y emisiones de tipo industrial y zonas urbanas y emisiones de tipo tráfico.

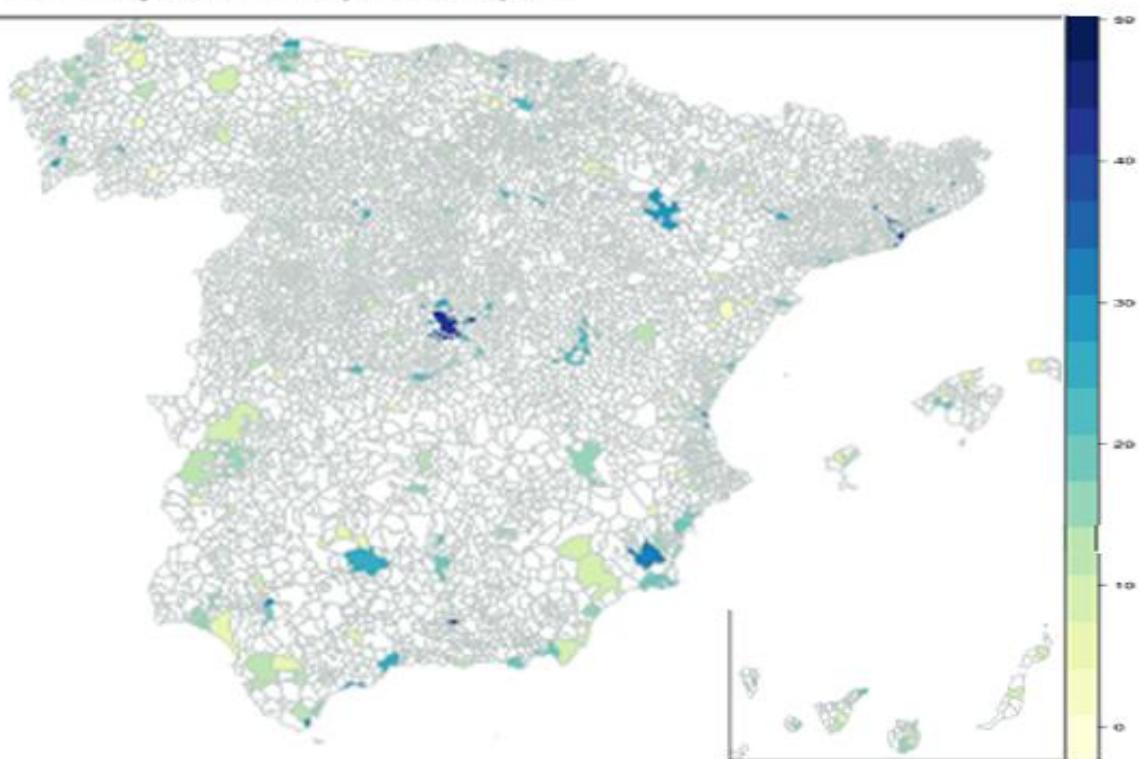
4.7. Gráficos de niveles de contaminación por municipios

Una buena muestra de cómo se distribuye el nivel de contaminantes en España es visualizar la información geográficamente por municipios sobre el propio mapa, incluidas las islas. A continuación, se muestra el estudio de los contaminantes que presentan más mediciones:

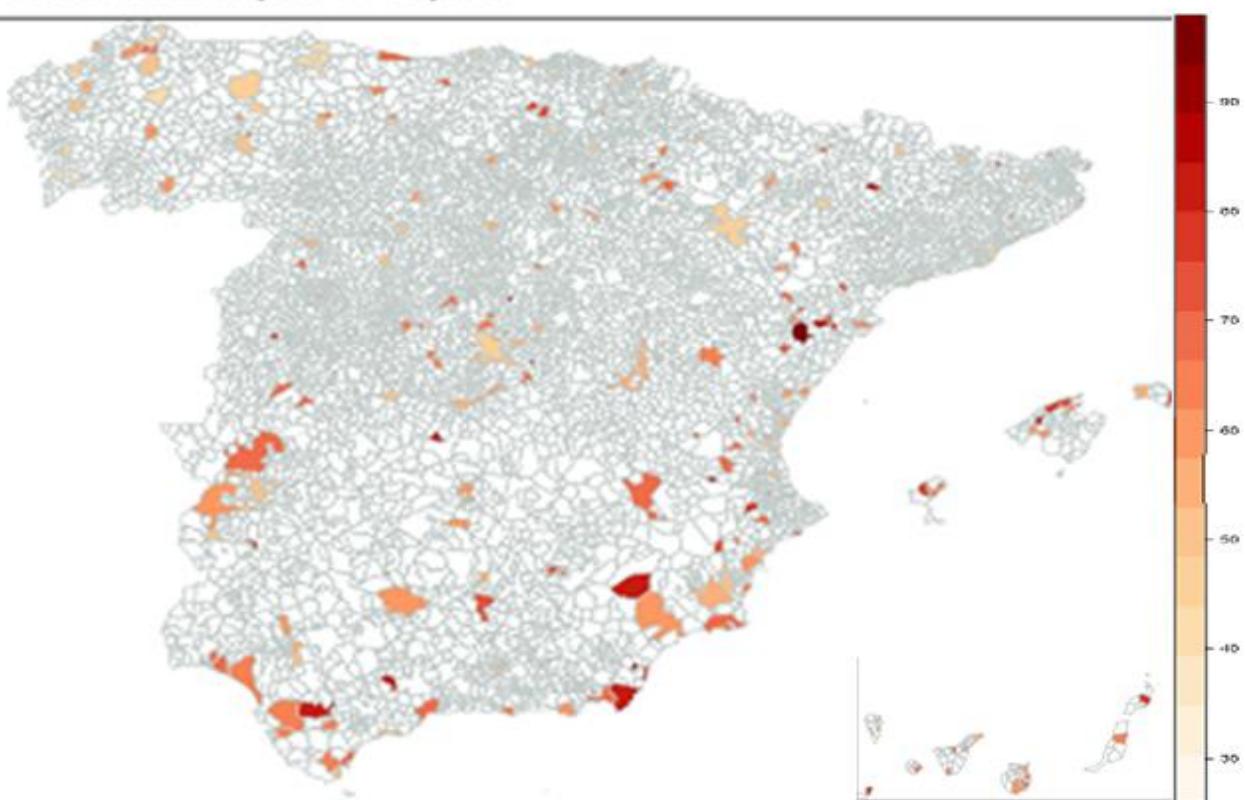
- *Dióxido de Nitrógeno*
- *Ozono*
- *Partículas en suspensión < 10µM*
- *Dióxido de Azufre*

En el *Anexo III* de este documento se incluirán los gráficos del resto de contaminantes, aunque todos los comentarios sobre cada contaminante estarán indicados en esta sección.

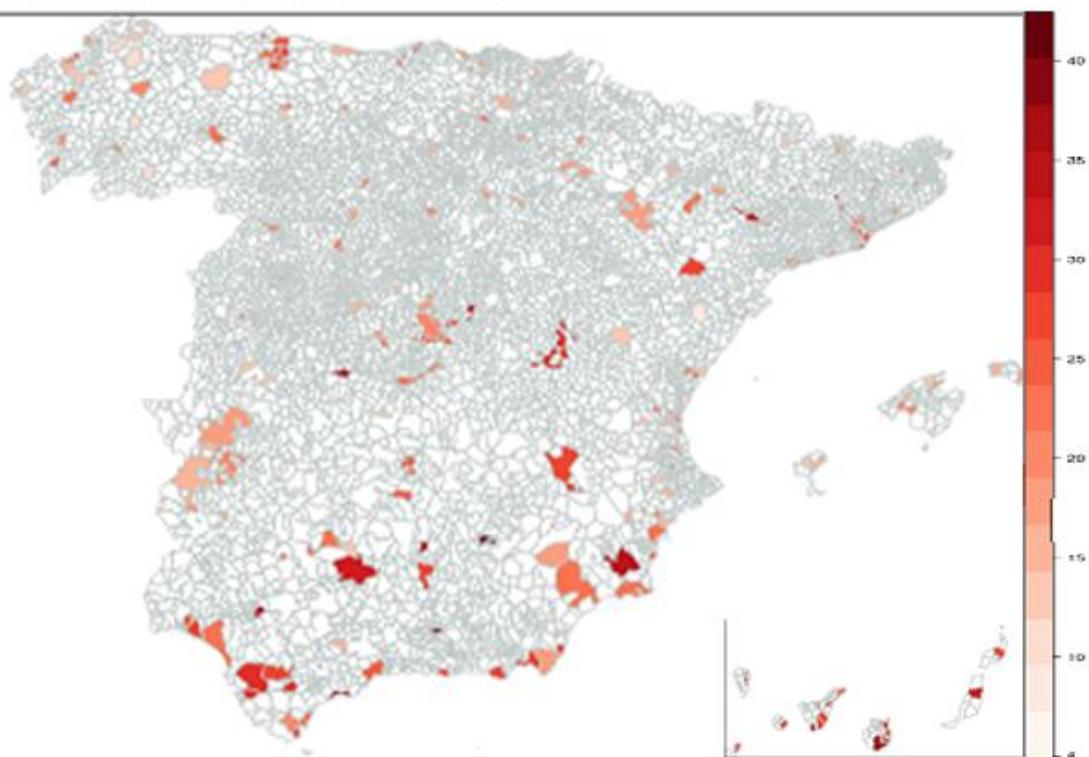
Dióxido de Nitrógeno x Municipios de España



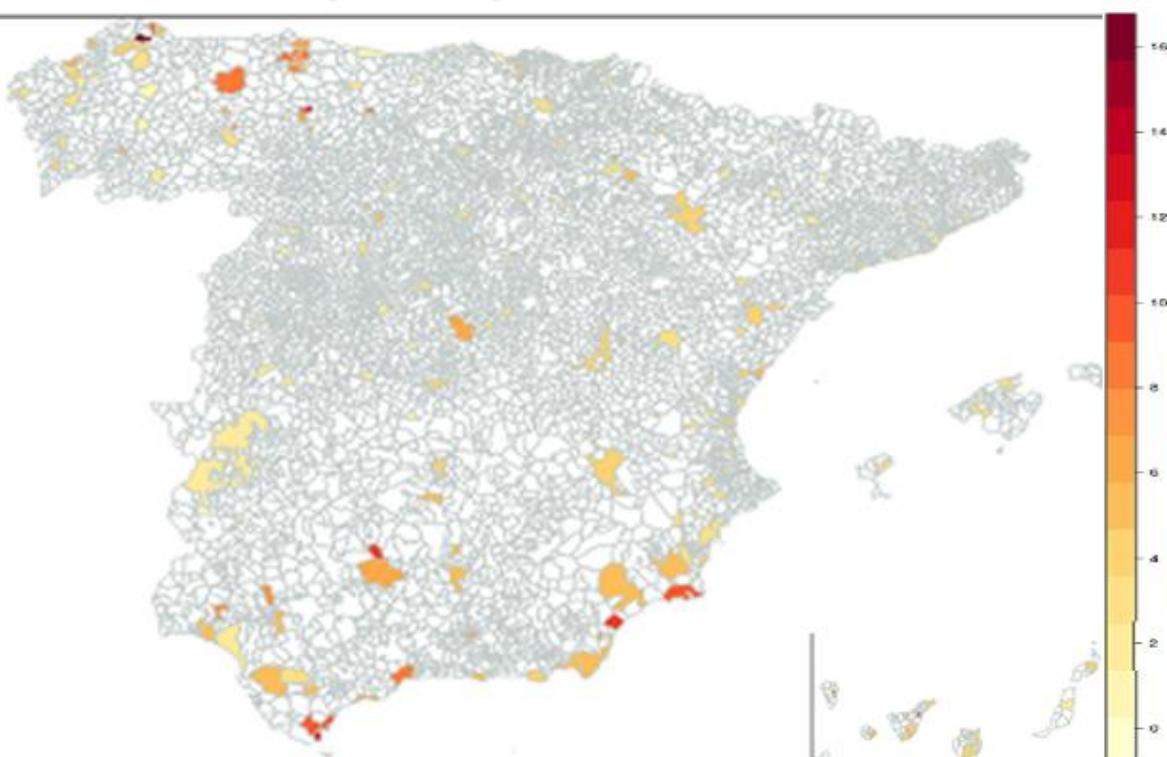
Ozono x Municipios de España



Partículas en Suspensión 10 μ M x Municipios de España



Dióxido de Azufre x Municipios de España



Lo que los gráficos demuestran, si observamos los 4 primeros mapas junto con el resto de los mapas del Anexo III de este documento, es que la cantidad de mediciones según distribución geográfica por municipios es muy pequeña. Hay muchísimas áreas que no presentan ningún tipo de medición, bien por falta de estaciones, bien por la imposibilidad de las estaciones o de los equipos de medición para capturar determinadas sustancias.

Arsénico

La información de este contaminante es bastante escasa. Destacan Huelva, Albacete y Lugo. Está relacionado con la proximidad a centrales térmicas del carbón, incineradoras, hornos de fundición y, en general, a industrias de este ámbito.

Benceno

La información de este contaminante es bastante escasa. Destacan los altos niveles muy localizados en Córdoba, Jaén y Asturias. En Córdoba existe un parque joyero que puede influir en la contaminación por esta sustancia.

Monóxido – Dióxido - Óxidos de Nitrógeno

Destacan los altos niveles de concentración de la comunidad de Madrid y, en menor medida, de Sevilla, Murcia, Zaragoza y Asturias. Queda evidente que, en el caso de Madrid, y otras ciudades grandes, se trata de emisiones provenientes del tráfico, mientras que en el caso de Asturias puede tratarse de la presencia de zonas mineras de carbón.

Partículas en Suspensión $<10\mu M$ y $< 2.5\mu M$

La presencia de partículas está bastante extendida en España, donde destacan las zonas de costa de Andalucía, Murcia, Córdoba, Extremadura, Asturias, Albacete, Cuenca e Islas Canarias. Sin duda una de las causas principales en la zona sur y las Islas Canarias es el polvo subsahariano.

Ozono

La contaminación por Ozono está bastante extendida por la península y las islas.

Plomo - Cadmio

La información de estos contaminantes es bastante escasa. Destacan muy por encima del resto Córdoba y Lugo.

Níquel

La información de este contaminante es bastante escasa. Destacan Cádiz, Málaga, Almería y Lugo.



Dióxido de Azufre

Las comunidades de Andalucía, Asturias, Extremadura, Galicia, Madrid y Murcia son donde mayores niveles de concentración existe y tiene un origen principalmente industrial.

4.8. Estudio sustancias más comunes

Tal y como se puede ver en los gráficos, existen sustancias que, por un lado, presentan mediciones de forma frecuente, mientras que, por el contrario, existen otras con una frecuencia de mediciones muy baja o nula. Las sustancias más frecuentes son:

- *Dióxido de Azufre*
- *Monóxido de Nitrógeno*
- *Dióxido de Nitrógeno*
- *Óxidos de Nitrógeno*
- *Partículas en Suspensión < 10 μM*
- *Ozono*
- *Arsénico*
- *Plomo*
- *Cadmio*
- *Níquel*

Las sustancias como el *Monóxido de Carbono*, *Partículas en Suspensión < 2.5 μM*, *Benzoapireno y el Benceno* presentan, en general, pocas mediciones.

Para poder hacer un estudio de clasificación y predicción del comportamiento de los contaminantes con mayor calidad, el estudio se va a centrar en aquellas sustancias que de forma más frecuente presentan mediciones

5 Análisis basado en agrupamiento usando K-Means

Una técnica que nos ayudará a entender mejor las relaciones entre los datos con los que estamos trabajando es aplicar un método de agrupamiento, el denominado método *K-Means*.

Este agrupamiento nos permitirá conocer cómo de importantes son unas sustancias respecto a la relación que tienen con otras sustancias y qué patrones de comportamiento tienen entre ellas, estableciendo diferentes grupos donde los valores de los atributos de los miembros de un mismo grupo serán o presentarán un comportamiento similar, por lo tanto, los miembros de grupos diferentes estarán influenciados por valores diferentes.



5.1. Estudios a realizar

Dado que tenemos múltiples alternativas para analizar los datos, tales como, ubicación geográfica, periodicidad, etc., elegimos las siguientes opciones de estudio para observar el comportamiento de los contaminantes:

- Agrupamiento por provincia, municipio, sustancia y valor de contaminación trimestral
- Agrupamiento por provincia, municipio y valor de contaminación trimestral de gases. Como gases se entiende el estudio específico de las siguientes sustancias:
 - *Dióxido de Azufre*
 - *Monóxido de Nitrógeno*
 - *Dióxido de Nitrógeno*
 - *Óxidos de Nitrógeno*
 - *Partículas en Suspensión* $<10\mu M$
 - *Ozono*
- Agrupamiento por provincia, municipio y valor de contaminación trimestral de metales. Como metales se entiende el estudio específico de las siguientes sustancias:
 - *Arsénico (*)*
 - *Plomo*
 - *Cadmio*
 - *Níquel*

(*) El Arsénico también es considerado un metal

5.2. Construcción del conjunto de datos

Para poder aplicar agrupamiento *K-Means* a nuestro estudio es necesario realizar una serie de transformaciones en nuestro conjunto de datos.

Dado que el *Benceno* es una de las sustancias contaminantes excluidas de este análisis de agrupamiento, siendo los *datos irregulares* el único conjunto de datos que lo incluyen, es conveniente indicar que los *datos irregulares* quedan fuera del alcance de este análisis.

Agrupación en períodos trimestrales

El sentido de tomar como periodo de referencia el trimestre y no cualquier otro, es que nuestro objetivo es intentar captar patrones de comportamiento de los contaminantes con un componente estacional, es decir, relacionado con la estación en la que estamos, primavera, verano, otoño o invierno.



Para ello se transforman los períodos de las mediciones del conjunto de datos diarios y horarios a períodos trimestrales:

Datos Diarios	Datos Horarios
Provincia	Provincia
Municipio	Municipio
Sustancia	Sustancia
Media Trimestre 1 (Día 1 ... Día 31)	Media Trimestre 1 (Hora 0 ... Hora 23)
Media Trimestre 2 (Día 1 ... Día 31)	Media Trimestre 2 (Hora 0 ... Hora 23)
Media Trimestre 3 (Día 1 ... Día 31)	Media Trimestre 3 (Hora 0 ... Hora 23)
Media Trimestre 4 (Día 1 ... Día 31)	Media Trimestre 4 (Hora 0 ... Hora 23)



Conjunto Único
Provincia
Municipio
Sustancia
Media Trimestre 1 [media datos diarios trimestre 1, media datos horarios trimestre 1]
Media Trimestre 2 [media datos diarios trimestre 2, media datos horarios trimestre 2]
Media Trimestre 3 [media datos diarios trimestre 3, media datos horarios trimestre 3]
Media Trimestre 4 [media datos diarios trimestre 4, media datos horarios trimestre 4]

Disposición de las mediciones de todas las sustancias por columnas

Se transforma la estructura del conjunto de datos pasando las mediciones por sustancia de fila a columna, tal y como se muestra en la siguiente tabla de ejemplo y con los períodos ya divididos por trimestres (T1 ... T4).

	Dióxido de Azufre_T1	Dióxido de Azufre_T2	...	Níquel_T4
Elemento 1	X1.0X	X2.0X	...	Xn.0X
Elemento 2	Y1.0Y	Y2.0Y	...	Yn.0Y
...
Elemento N	Z1.0Z	Z2.0Z		Zn.0Z

Por *elemento* o *muestra* se entienden los valores de todas las sustancias contaminantes de todos los cuatrimestres de una fila de todo nuestro conjunto de datos objeto de estudio.

Eliminar Datos Faltantes

En el estudio se han encontrado una gran cantidad de datos faltantes, a pesar de trabajar con valores medios anuales. El dato faltante o nulo no computa para calcular cualquier valor medio.

Para los valores faltantes existentes se ha aplicado una técnica adicional consistente en *imputar* para esa misma sustancia su valor correspondiente del año 2016, siempre que tenga un valor informado. De esta forma, es mejor tener un valor medio del año anterior que no tener nada.

Convertir datos a valores numéricos

Debido a que existen diferentes medidas de distancia para trabajar con datos numéricos, es recomendable, siempre que sea posible, transformar valores categóricos o no numéricos a valores numéricos, además de que los algoritmos matemáticos funcionarán siempre mejor. En nuestro caso esta técnica ha sido aplicada.

Marca Coche	Color
Ford	Azul
Toyota	Gris
Seat	Rojo

Valor Color
1
2
3

Normalizar datos

La *normalización* de datos permite ajustar y agrupar los valores numéricos medidos en diferentes escalas a una escala común, de forma que no existan valores muy grandes o muy pequeños, que tengan demasiado o escaso peso en el método de clasificación. En este caso aplicaremos una normalización de datos con valores entre 0 y 1.

Artículo	Importe
Casa	1000000
TV	100
Moto	5995

Importe Normalizado
1
0
0.00589559

La normalización de datos no es más que aplicar la siguiente fórmula sobre el valor a normalizar:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

donde:

- X' es el valor obtenido normalizado
- X es el valor a normalizar
- X_{min} es el valor mínimo del conjunto de datos a normalizar
- X_{max} es el valor máximo del conjunto de datos a normalizar

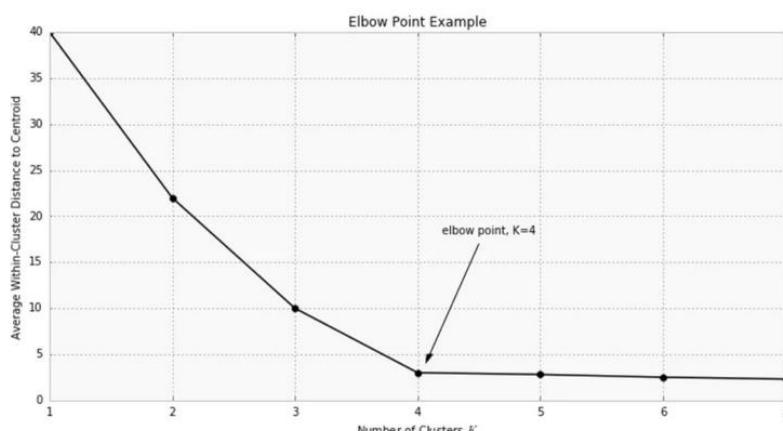


5.3. Técnicas para calcular el valor de K óptimo

Una vez transformados los datos siguiendo los pasos indicados, nos disponemos a ejecutar el agrupamiento usando *K-Means* por trimestre para todas las sustancias, para cada estudio indicado en el apartado *5.1 Estudios a realizar*. Para ello debemos calcular la *K* óptima previamente. A continuación, se indican los resultados obtenidos aplicando las técnicas introducidas en el apartado *2.3 Técnicas de aprendizaje no supervisado*.

Método Elbow

Una de las métricas usadas para comparar resultados es la *distancia media entre los puntos de datos y su centroide*. Como el valor de la media disminuirá a medida que aumentemos el valor de *K* (puesto que a mayor a mayor número de centroides menor es el diámetro de los grupos de puntos asignados a cada uno de ellos), debemos utilizar la distancia media al centroide en función de *K* y encontrar un punto de inflexión tal, que aunque aumentemos el valor de *K*, la distancia media apenas sufrirá variaciones, es decir, si representamos gráficamente la distancia media en función de *K*, dicho punto viene determinado porque la tasa de descenso se vuelve más constante, más afilada, es lo que se denomina *Método Elbow* o *método del codo*. Aquí vemos un ejemplo:



Análisis de Silueta

Está basado en la combinación de la *cohesión* del *clúster*, cómo de cerca están los puntos en un *clúster* entre ellos, y la *separación* entre clústers, cómo de alejados están los clústers entre ellos, por lo tanto, mide cuán distante está un punto de los otros grupos. La aplicación de esta técnica devuelve un valor numérico o coeficiente comprendido entre -1 y 1.

- Un valor cercano a 1 indica que el punto está lejos de los otros grupos, por lo tanto, ha sido asignado al grupo correcto
- Un valor alrededor de 0 indica que el punto está cerca de otros grupos
- Un valor cercano a -1 indica que el punto está situado en el grupo equivocado



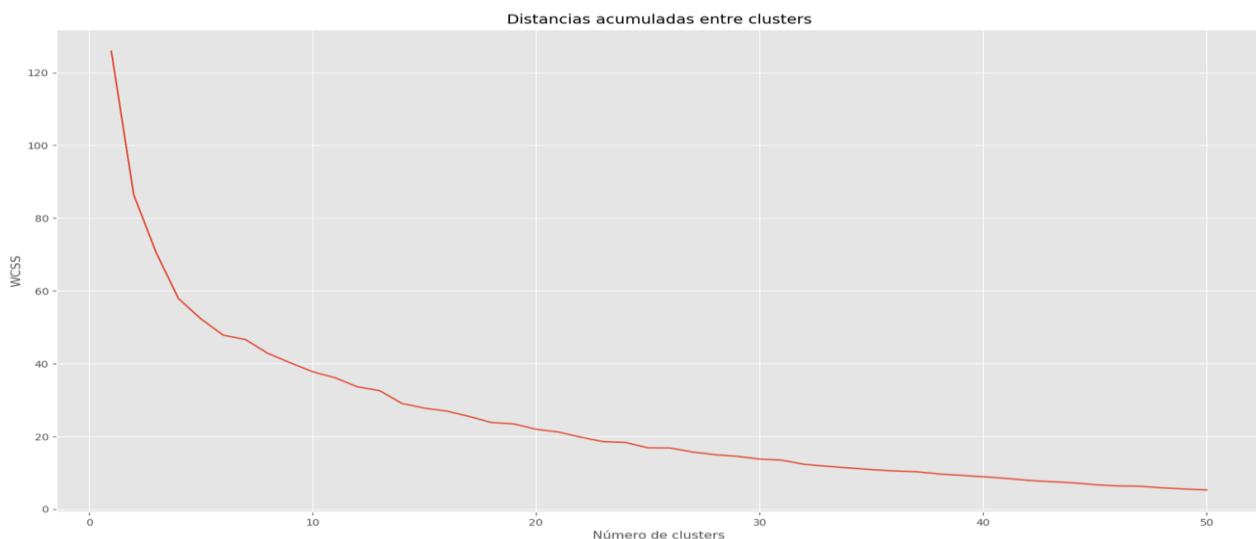
El valor de K será aquel cuyo coeficiente obtenido tenga un valor alto.

Se mostrarán representaciones gráficas de los resultados de este método en nuestro caso de estudio.

Agrupamiento por Provincia, Municipio, Contaminante y Trimestre

Método Elbow

En la representación gráfica de los valores generados por esta técnica se debería apreciar un cambio brusco en la evolución de la línea representando una forma similar a la de un brazo y su codo. El punto en el que se observa ese cambio brusco nos debería decir el número óptimo de *clústers* a seleccionar para ese conjunto de datos. Tras la ejecución del método, se muestra como resultado:

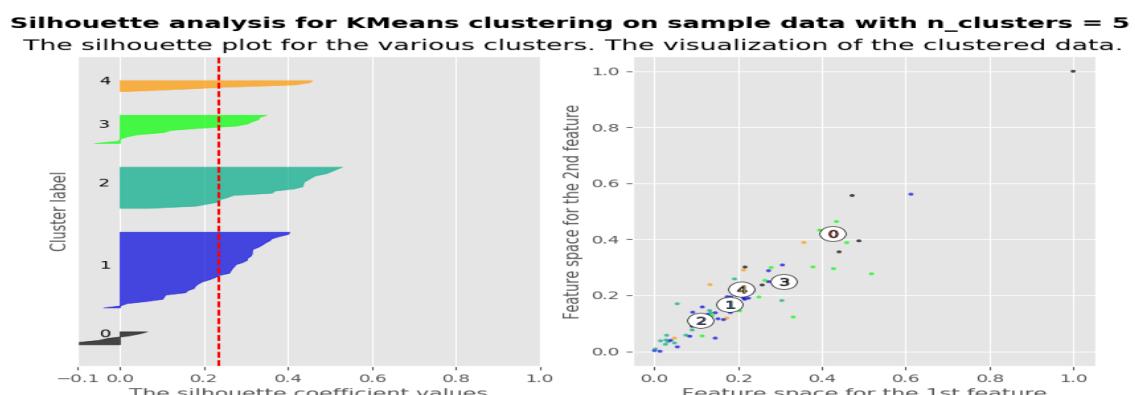
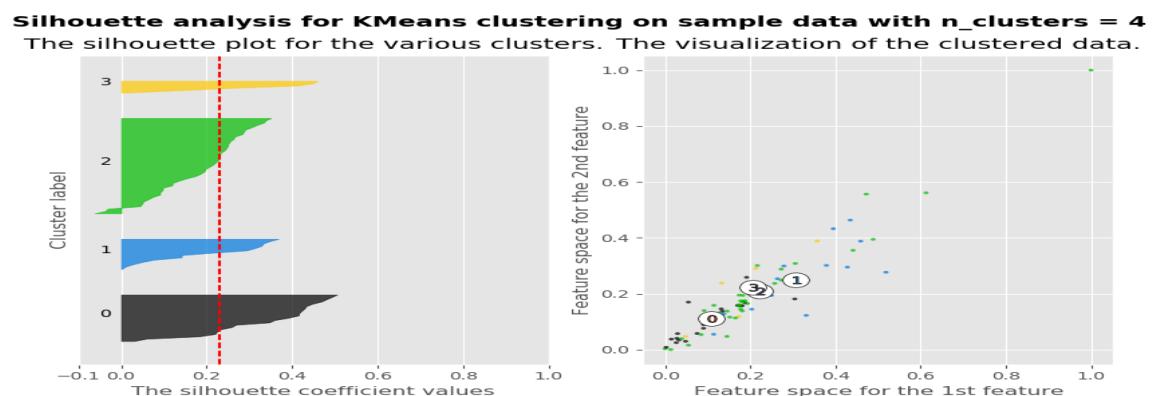
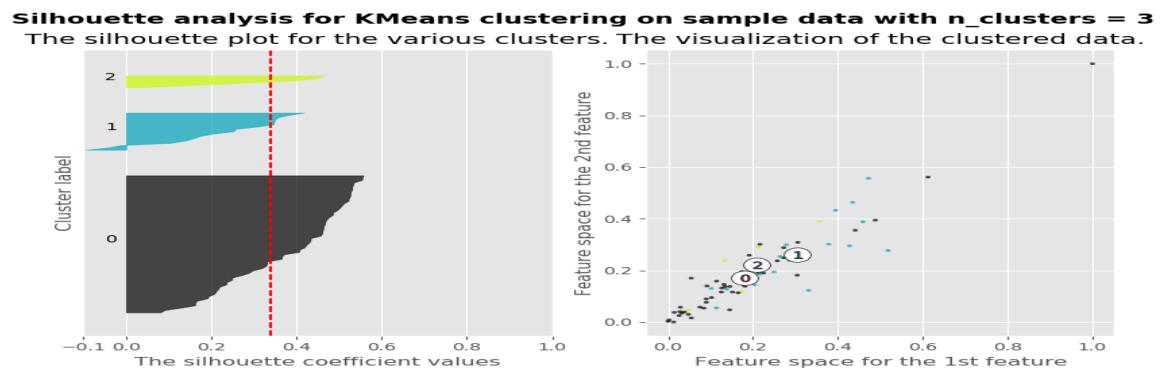
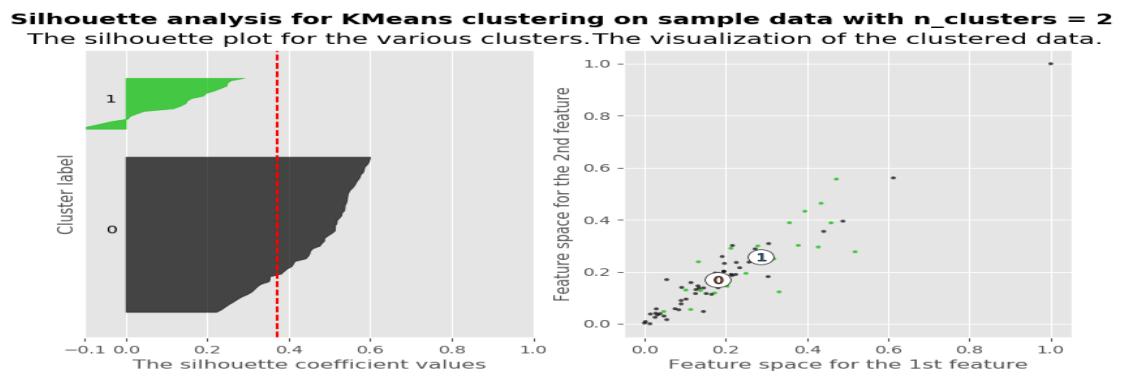


No queda claramente identificado el número óptimo de K , pudiendo oscilar su valor entre 5 y 10 *clústers*, lo que hace necesario contrastar este resultado aplicando otras técnicas adicionales para poder establecer una comparativa.

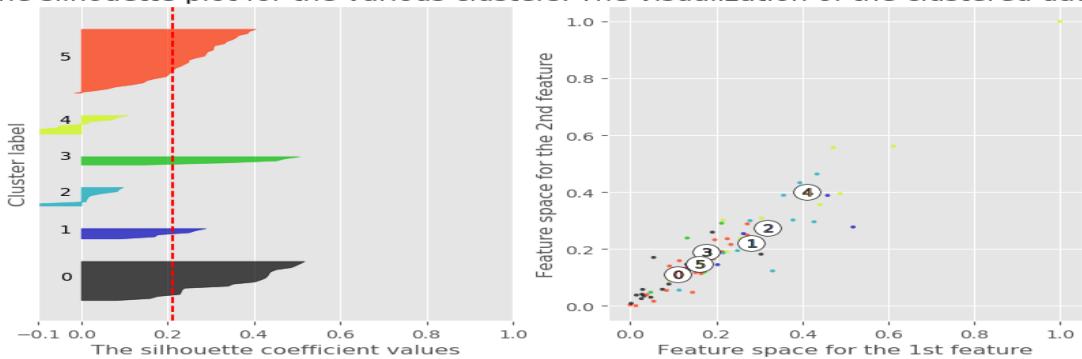
Análisis de Silueta

Esta técnica es necesario aplicarla para varios Ks, de esta forma se podrán comparar los resultados entre sí, quedándonos con el óptimo.

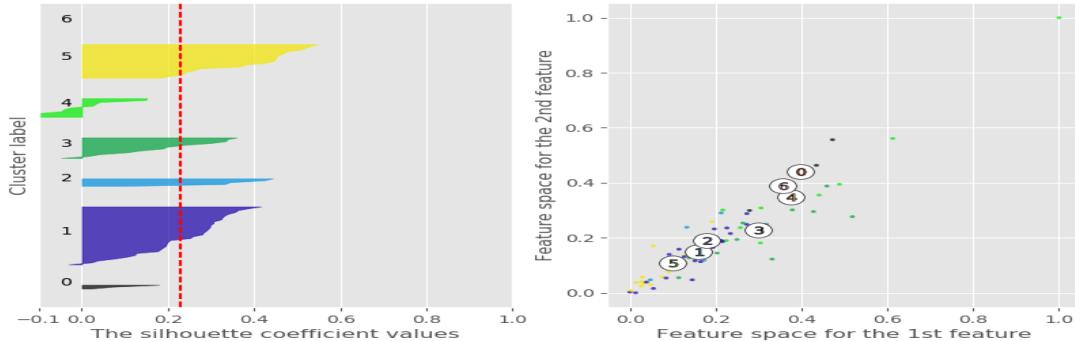




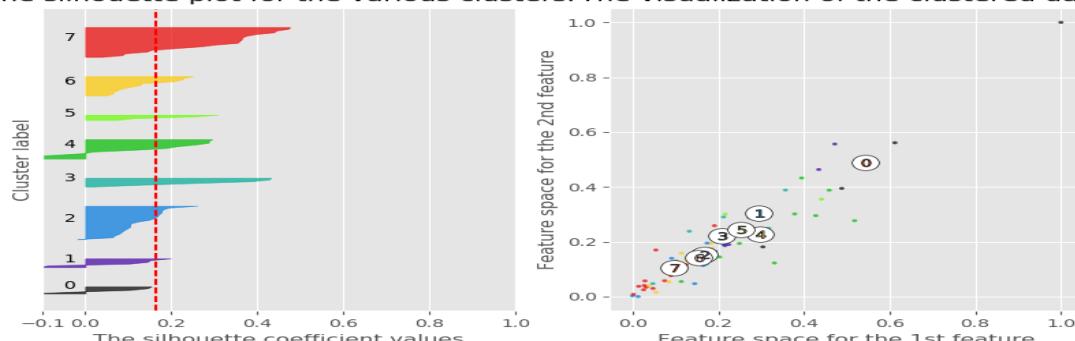
Silhouette analysis for KMeans clustering on sample data with n_clusters = 6
The silhouette plot for the various clusters. The visualization of the clustered data.



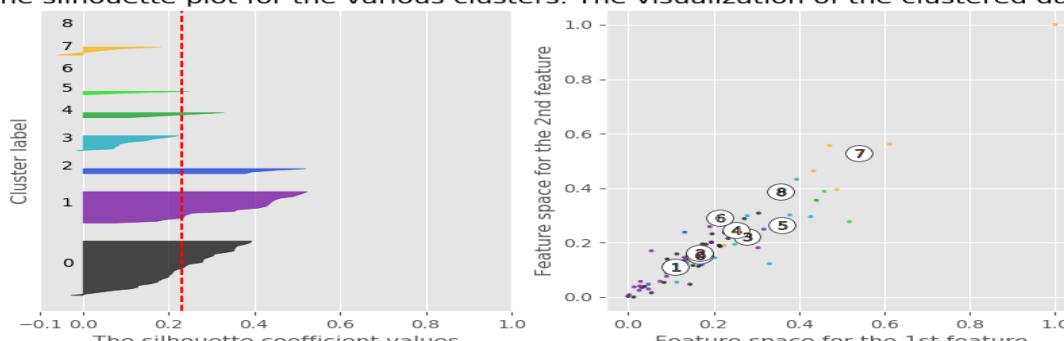
Silhouette analysis for KMeans clustering on sample data with n_clusters = 7
The silhouette plot for the various clusters. The visualization of the clustered data.

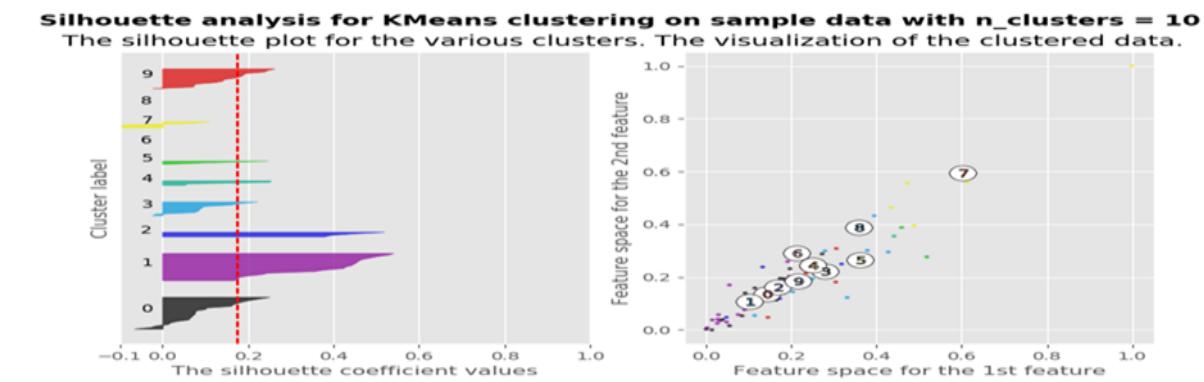


Silhouette analysis for KMeans clustering on sample data with n_clusters = 8
The silhouette plot for the various clusters. The visualization of the clustered data.



Silhouette analysis for KMeans clustering on sample data with n_clusters = 9
The silhouette plot for the various clusters. The visualization of the clustered data.





Según los gráficos, la imagen de la izquierda muestra la pertenencia de una sustancia a un grupo determinado, siendo el total de grupos el número de *clusters* indicado.

La línea roja es la puntuación media para el grupo en consideración.

Para que este sea un buen valor para el número de clúster, uno debe considerar los siguientes puntos:

- El valor medio debe ser lo más cercano posible a 1
- La gráfica de cada grupo debe estar por encima del valor medio tanto como sea posible. Cualquier región del gráfico por debajo del valor medio no es deseable.
- Por último, el ancho de la trama debe ser lo más uniforme posible.

La imagen de la derecha es la visualización de la asignación del clúster.

Además de los gráficos, la herramienta te proporciona un coeficiente que mide el grado de cohesión de los datos:

Número de Clústers (K)	Coeficiente
2	0.3718625770
3	0.3401360053
4	0.2299247310
5	0.2356788707
6	0.2129378534
7	0.2273399160
8	0.1652227403
9	0.2327897132
10	0.1749108367

La interpretación de con qué K debemos quedarnos es una decisión basada en los valores aportados por el coeficiente de silueta y los resultados gráficos.

Según la representación gráfica y el valor del coeficiente que mide el grado de cohesión, nuestra K óptima podría ser un valor entre 2 y 5.



Dentro de ese rango de valores, bien es cierto que si trabajamos con valores bajos de K ($K = 2, K = 3$), estaremos haciendo agrupaciones con poca variabilidad de datos, quedando un agrupamiento ciertamente *pobre*, mientras que si utilizamos una K con valores altos ($K = 4, K = 5$) también es posible que perdamos algo de precisión, ya que un elemento puede estar siendo asignado a un grupo que no le corresponde.

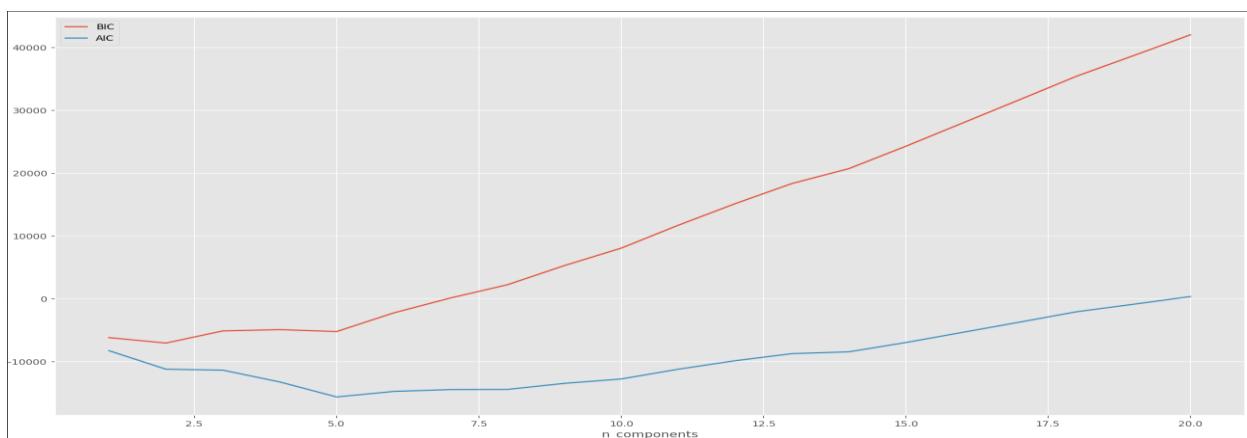
Valorando estas opciones, es el criterio del analista, junto con la aportación de resultados de otras técnicas adicionales, el que tiene que tomar la decisión de cuál debe ser la K óptima.

En nuestro caso tiene más sentido quedarnos con una $K = 4$ o $K = 5$, a pesar de que se pueda perder algo de precisión, pero por otra parte también se gana en la identificación de más grupos.

Gaussian Mixture Models

Para aplicar esta técnica, tras ejecutar el algoritmo con un máximo de 20 componentes, simplemente hay que mostrar gráficamente el resultado de las métricas:

- AIC: Akaike Information Criterion
- BIC: Bayesian Information Criterion



El número óptimo de *clusters* K será el valor que minimiza el *AIC* o el *BIC*, parece que ambas métricas convergen en un $K = 5$.

Elección de un K Óptimo según las diferentes técnicas aplicadas

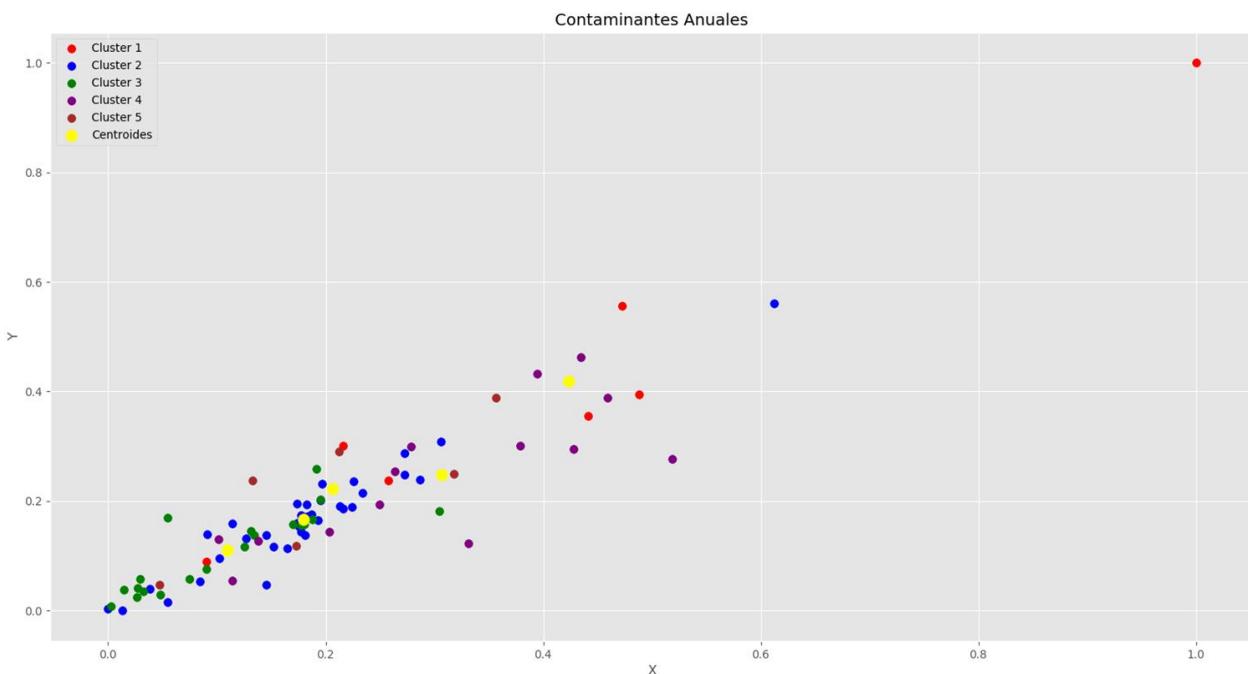
En resumen, los resultados obtenidos para cada técnica de cálculo del valor óptimo de K son:

Método	K Óptima
Método Elbow	Oscila entre K = 5 y K= 10
Análisis de Silueta	K = 4 K = 5
Gaussian Mixture Model	K = 5

Si tenemos en cuenta en conjunto todos los resultados, parece que todo converge a un valor óptimo de $K = 5$

Ejecución de K-Means por Provincia, Municipio, Contaminante y Trimestre

Una vez calculado el valor de K inicial, procedemos a ejecutar el algoritmo con el siguiente resultado:



El gráfico muestra el resultado final del algoritmo. Durante su ejecución se han hecho los siguientes pasos:

- Asignación de elementos a los centroides
- Actualización de centroides

Según se ha explicado en el apartado *2.4 Técnicas de aprendizaje no supervisado*, esta técnica identifica agrupaciones en un espacio d-dimensional.

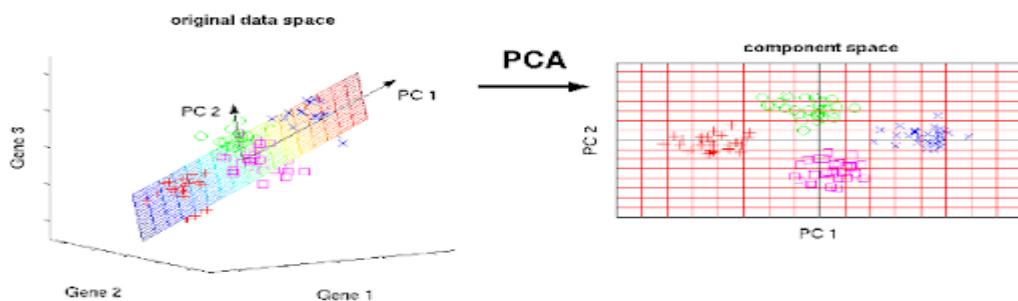
Como se aprecia en el gráfico, existen elementos de diferentes grupos que se superponen, esto es debido a que el gráfico muestra la información en 2 dimensiones, cuando realmente los elementos se distribuyen en un espacio de más de dos dimensiones.

Reducción de Dimensionalidad. PCA

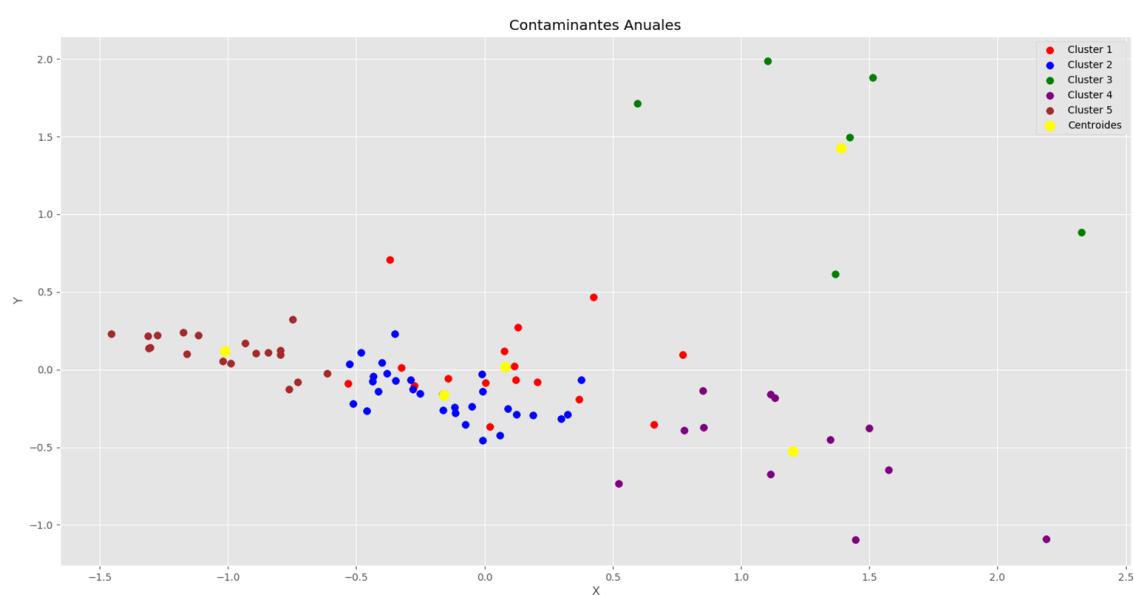
Una herramienta utilizada para poder representar gráficamente elementos de un espacio d-dimensional en un espacio bidimensional es la denominada técnica de *Análisis de Componentes Principales*, conocido por las siglas *PCA*.

Esta técnica reduce la dimensionalidad del espacio de entrada aplicando una transformación en los m atributos originales en otro conjunto de atributos p donde $p \leq m$.

Una forma de representar esta transformación gráficamente podría ser:



Si aplicamos *PCA* para generar el gráfico de *K-Means* reduciendo las dimensiones del espacio de las muestras obtenemos:



Con este gráfico se ve de forma más clara cómo se distribuyen las muestras entre los diferentes grupos o centroides.

Junto con el gráfico, es importante incluir también el porcentaje de información explicado por cada uno de los componentes seleccionados.



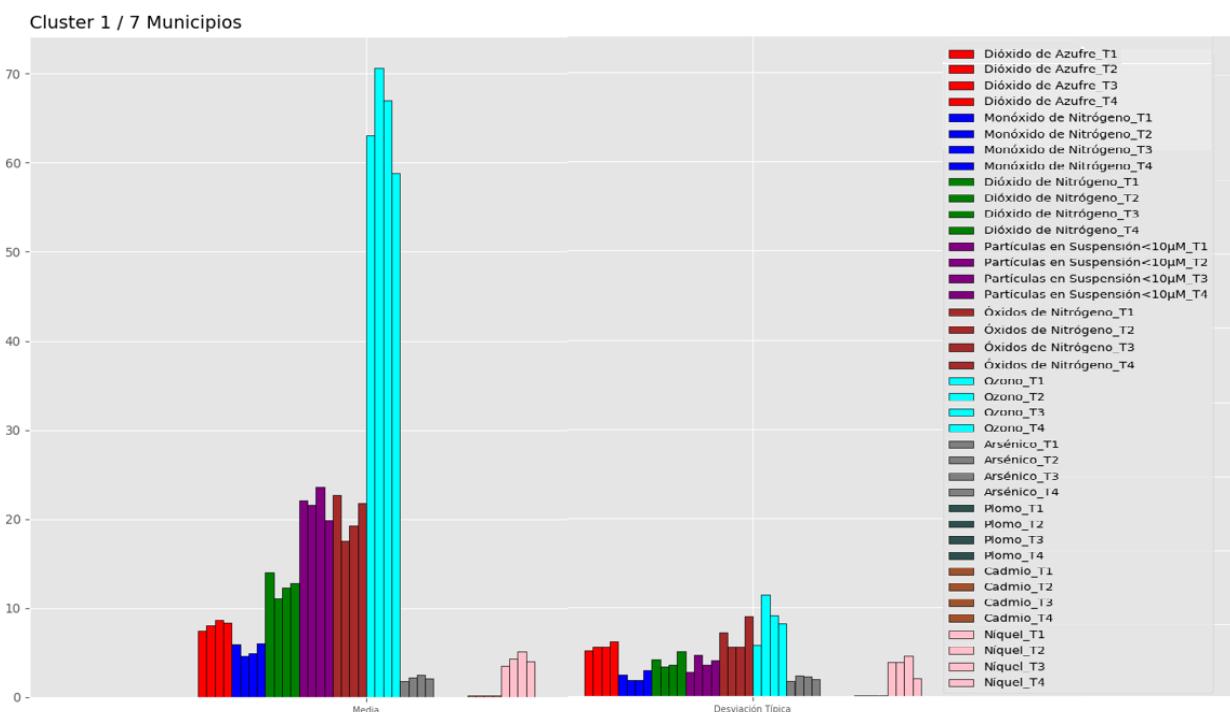
Ratio de información capturada para los 5 componentes respectivamente:

46.18 %	17.06 %	7.8 %	4.91 %	4.66 %
---------	---------	-------	--------	--------

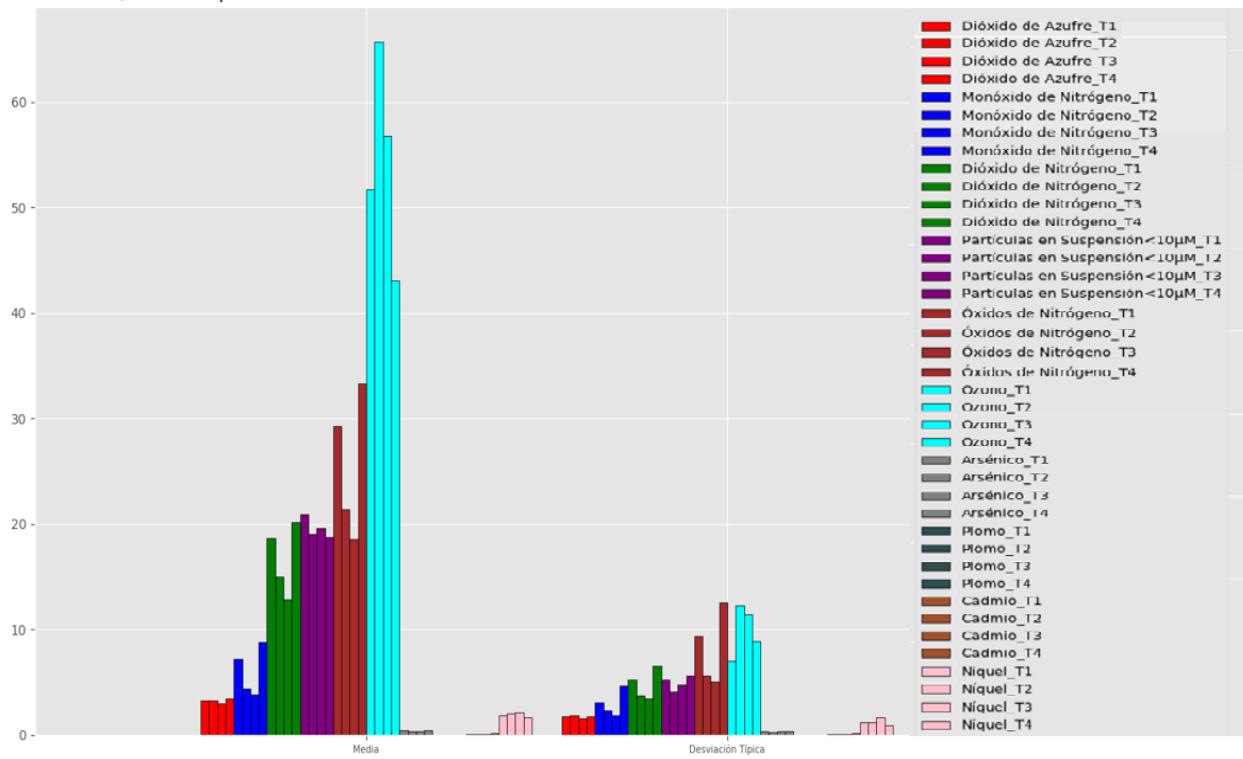
Es importante destacar que esta técnica solamente se utiliza para representar gráficamente la información reduciendo las dimensiones, pero no sirve para utilizarlo como análisis porque está eliminando información del conjunto de datos.

Visualización de los Grupos

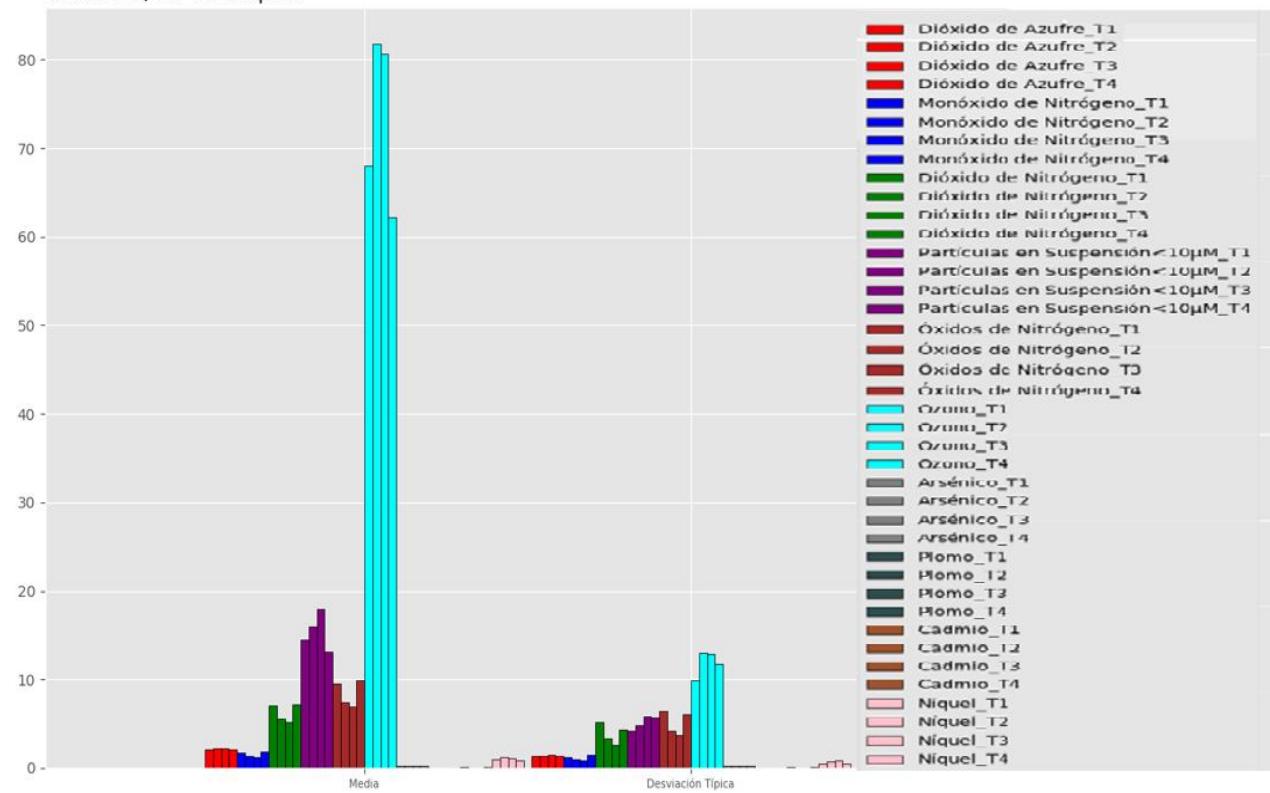
Una vez ejecutado el método *K-Means*, vamos a visualizar los elementos que componen cada grupo o clúster identificado. Los valores de los contaminantes corresponden a la media y a la desviación típica de los elementos.

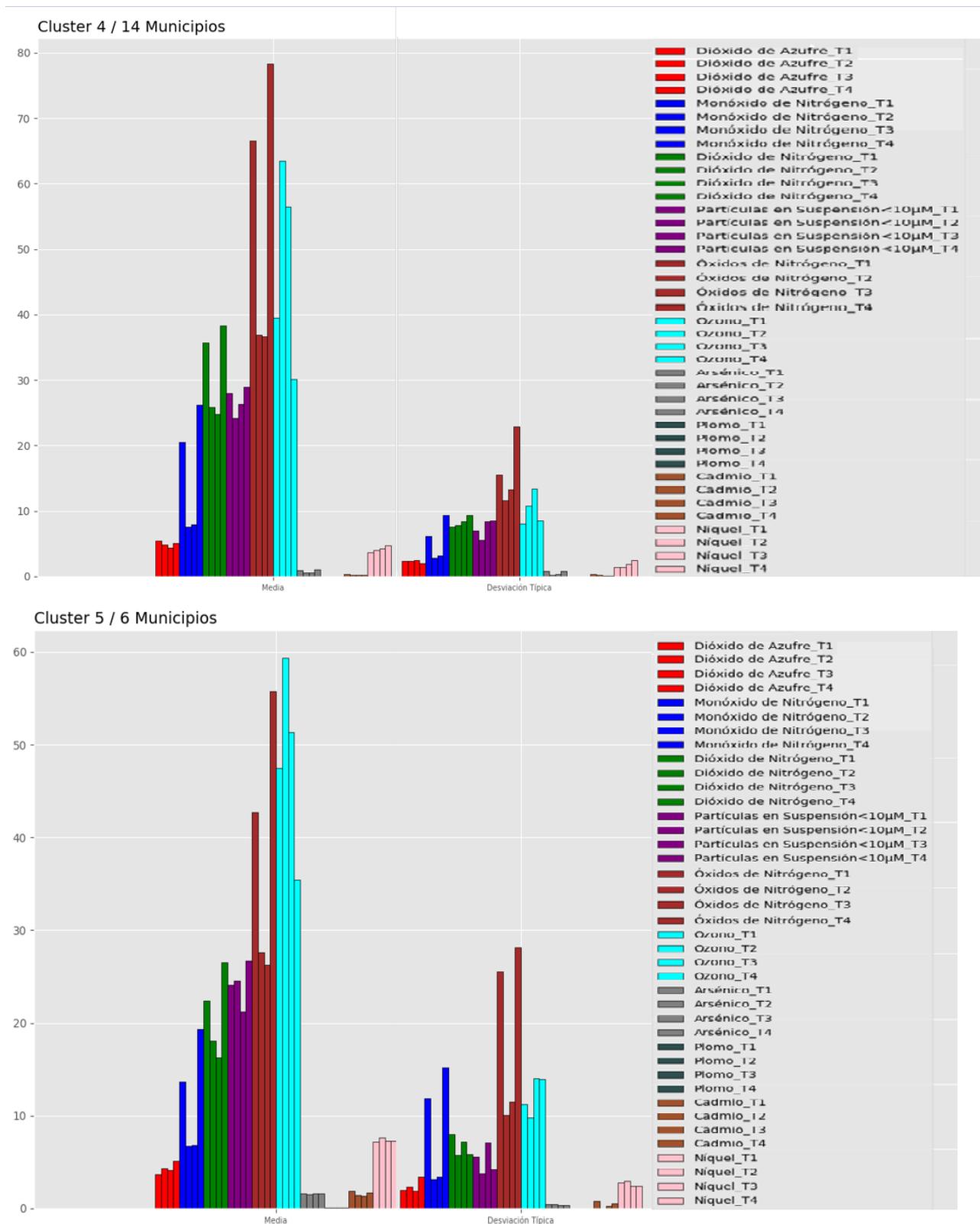


Cluster 2 / 36 Municipios



Cluster 3 / 20 Municipios





Análisis de la Información

Clúster 1

El criterio de agrupación corresponde a los valores más altos de *Dióxido de Azufre* y *Arsénico*. También presenta altos niveles de contaminación de *Níquel* y de *Ozono*.

Los municipios de esta agrupación son:

Municipio	Provincia
Ibiza	Ibiza
Mahón	Menorca
San Fernando	Cádiz
San Roque	Cádiz
Moguer	Huelva
Huelva	Huelva
La Robla	León

El resto de los contaminantes se han agrupado por presentar valores similares dentro de un mismo rango.

Llama la atención, que las moléculas de *Nitrógeno* (*Monóxido*, *Dióxido* y *Óxido*) presentan valores más altos en los trimestres 1 y 4.

En el caso de Cádiz y Huelva, tras analizar posibles motivos y centrando el estudio en la contaminación de *Arsénico* y *Dióxido de Azufre*, está relacionado con la presencia de un recinto de empresas químicas y de industria relacionada con la fundición y producción de cobre.

Clúster 2

No destaca ningún valor de contaminante que predomine sobre el resto, los contaminantes, en conjunto, se han agrupado por valores similares y posiblemente este sea el criterio de agrupación, aunque sí cabe destacar que los valores son medios, ya que no corresponden a los valores más altos ni a los más bajos del conjunto de todos los *clústers*.

Aun así, la agrupación sí presenta valores elevados de *Ozono*.

Es el grupo con más elementos y geográficamente está muy repartido por España.

Los municipios de esta agrupación son:

Municipio	Provincia	Municipio	Provincia
Alicante	Alicante	Castro Urdiales	Cantabria
Elche	Alicante	Corrales de Buelna	Cantabria
Torrevieja	Alicante	Reinosa	Cantabria



Almería	Almería	Santander	Cantabria
Mérida	Badajoz	Torrelavega	Cantabria
Palma de Mallorca	Palma de Mallorca	Segovia	Segovia
Berga	Barcelona	Vila-Seca	Tarragona
Igualada	Barcelona	Toledo	Toledo
Aranda de Duero	Burgos	Alzira	Valencia
Miranda de Ebro	Burgos	Burjassot	Valencia
Alcora	Castellón	Gandía	Valencia
Onda	Castellón	Paterna	Valencia
Puertollano	Ciudad Real	Sagunto	Valencia
Ponferrada	León	Torrebaja	Valencia
Cartagena	Murcia	Torrent	Valencia
Pamplona	Navarra	Medina del Campo	Valladolid
Tenerife	Tenerife	Alagón	Zaragoza

Las moléculas de *Nitrógeno* presentan valores más altos en los trimestres 1 y 4.

Clúster 3

Presenta los valores más altos de *Ozono*.

Por contra, coincide con los valores más bajos de la molécula de *Nitrógeno*, lo que indica una relación a la inversa, a mayor nivel de *Ozono*, menor nivel de *Nitrógeno*.

Los valores de *Ozono* son mayores en los trimestres 2 y 3, lógicamente por coincidir con periodo de primavera y verano.

Geográficamente está distribuido por España, aunque está más localizado en ciudades que no son grandes y de interior. Los municipios de esta agrupación son:

Municipio	Provincia	Municipio	Provincia
Albacete	Albacete	Morella	Castellón
Alcoy	Alicante	San Jorge	Castellón
Pinoso	Alicante	Víznar	Granada
Badajoz	Badajoz	Campisábalos	Guadalajara
Zafra	Badajoz	Almonte	Huelva
Bunyola	Palma de Mallorca	Els Torms	Lleida
Cáceres	Cáceres	El Atazar	Madrid
Plasencia	Cáceres	Llanes	Asturias
Toril	Cáceres	San Nicolás del Puerto	Sevilla
Cirat	Castellón	San Pablo de los Montes	Toledo

El resto de los contaminantes se entiende que se han agrupado por valores similares.

Las moléculas de *Nitrógeno* presentan valores más altos en los trimestres 1 y 4.

Analizando los motivos y centrando el estudio en la contaminación de *Ozono*, está relacionado con:



- La existencia de más estaciones meteorológicas en poblaciones de tamaño medio fuera de núcleos urbanos grandes, de grandes ciudades.
- Las ciudades grandes ejercen un efecto de sustentación frente al *Ozono* por su arquitectura, que ayuda a retener a esta sustancia en capas más altas, por edificios, construcciones, etc.
- Adicionalmente, los municipios son de interior, no hay municipios de costa. Esto quiere decir que el efecto del *Ozono* es superior en zonas de interior respecto a zonas de costa.

Clúster 4

Presenta los valores más altos de contaminación de *Nitrógeno*, con una gran diferencia sobre el resto.

Los trimestres 1 y 4 son los de mayor contaminación.

También se observa presencia de *Arsénico* y también altos valores de *Níquel* y *Dióxido de Azufre*. Está relacionado con las emisiones producidas en las ciudades grandes, como consecuencia del tráfico, de presencia de zonas industriales, de mayor uso de calefacción, etc ya que destacan ciudades grandes.

Los municipios de esta agrupación son:

Municipio	Provincia	Municipio	Provincia
Barcelona	Barcelona	Móstoles	Madrid
Manlleu	Barcelona	Málaga	Málaga
Prat de Llobregat	Barcelona	Avilés	Asturias
Sant Vicent dels Horts	Barcelona	Gijón	Asturias
Granada	Granada	Sevilla	Sevilla
Bailén	Jaén	Valencia	Valencia
Madrid	Madrid	Bilbao	Vizcaya

También destaca que la contaminación por *Material Particulado* es la mayor, pareciendo estar relacionado por lo tanto con que parte de ese material particulado es generado también por las emisiones del tráfico, industria, etc.

Clúster 5

Los municipios de este clúster, se podría decir que, en general, son los que más contaminación presentan si valoramos en global contaminación de gases y metales.

Presencia de altos valores de *Ozono*, *Moléculas de Nitrógeno* y *Material Particulado*

Altos niveles de contaminación de metales: *Níquel*, *Cadmio* y *Arsénico*

Altos niveles de contaminación de *Dióxido de Azufre*

Coincide en su totalidad con municipios de Galicia y Córdoba.

Los municipios de esta agrupación son:



Municipio	Provincia
Córdoba	Córdoba
A Coruña	A Coruña
Santiago de Compostela	A Coruña
Lugo	Lugo
Ourense	Ourense
Vigo	Pontevedra

Analizando los motivos de esta contaminación, en el caso de Galicia, está relacionado con un desarrollo urbanístico caótico que hace que esta zona sea la zona más densa industrial de esta comunidad autónoma. Existe industria de fabricación y procesado de aluminio, carbón, hierros y aceros, madera, celulosa, colas, resinas y distribución de químicos, así como la existencia de refinerías y centrales térmicas.

En el caso de Córdoba, por un lado, destacan las elevadas concentraciones de *Ozono*, sobre todo en verano, debido al calor y a la falta de lluvia. El origen de la generación de ese *Ozono* está relacionado con la existencia de actividad industrial de centrales térmicas, fundiciones, depósitos agroquímicos, rellenos sanitarios, vertederos y canteras. En gran parte del territorio yacen metales pesados, sustancias radioactivas y otros residuos, así como la deforestación a la que también se ha visto sometido.

Conclusiones

Ha quedado manifiesto que la principal causa de la contaminación en España está motivada principalmente por:

- Tráfico
- Zonas industriales de producción de materia prima y centrales térmicas sin un plan urbanístico acorde con un desarrollo de tejido industrial

Los trimestres de mayor contaminación de *Nitrógeno* son los trimestres 1 y 4, periodo que coincide con otoño e invierno, es decir, mayor desplazamiento de tráfico rodado, producción industrial y consumo energético.

La contaminación por *Plomo* apenas tiene presencia en el estudio.

La relación entre el *Ozono* y el *Nitrógeno* también es evidente, cuanto mayor es el *Ozono*, menor es el *Nitrógeno* y viceversa.

También se observa que a mayores niveles de *Nitrógeno* existe mayor presencia de contaminación de *Material Particulado*, *Arsénico* y *Dióxido de Azufre*, por tanto, esta circunstancia va muy ligada al tráfico rodado.

A continuación, centraremos nuestro estudio únicamente en los contaminantes gaseosos



Agrupamiento por Provincia, Municipio, Gases y Trimestre

En este estudio se seguirán los mismos pasos indicados en el estudio anterior.

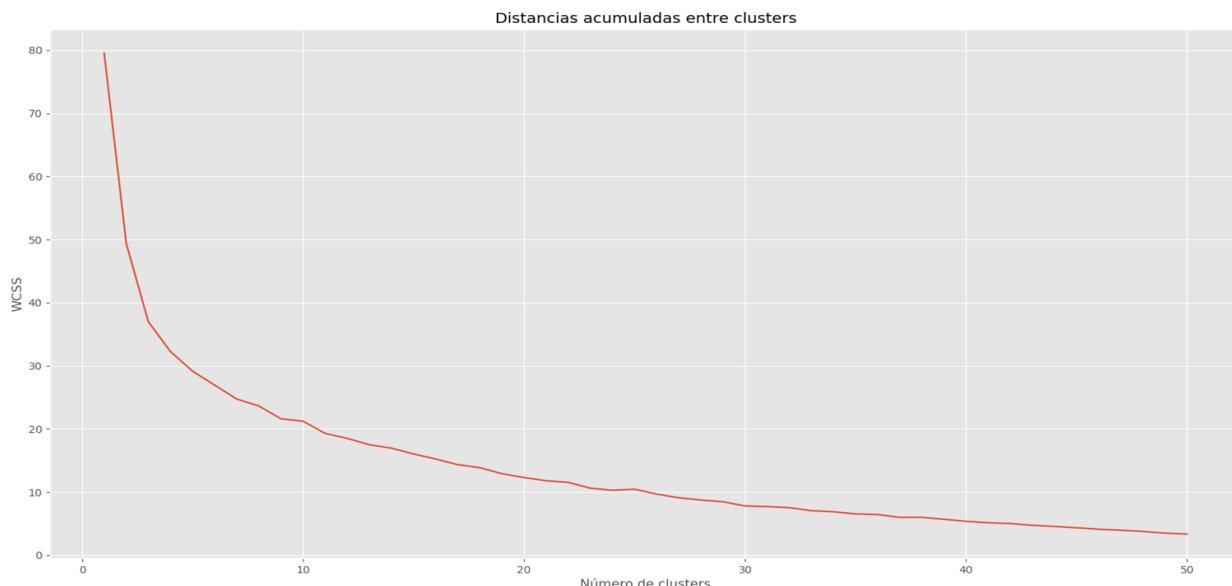
Utilizando exactamente las mismas técnicas, pasaremos a hacer un estudio específico de los siguientes contaminantes:

- *Dióxido de Azufre*
- *Monóxido de Nitrógeno*
- *Dióxido de Nitrógeno*
- *Óxidos de Nitrógeno*
- *Partículas en Suspensión < 10 μM*
- *Ozono*

El primer paso es determinar el valor óptimo del parámetro K.

Método Elbow

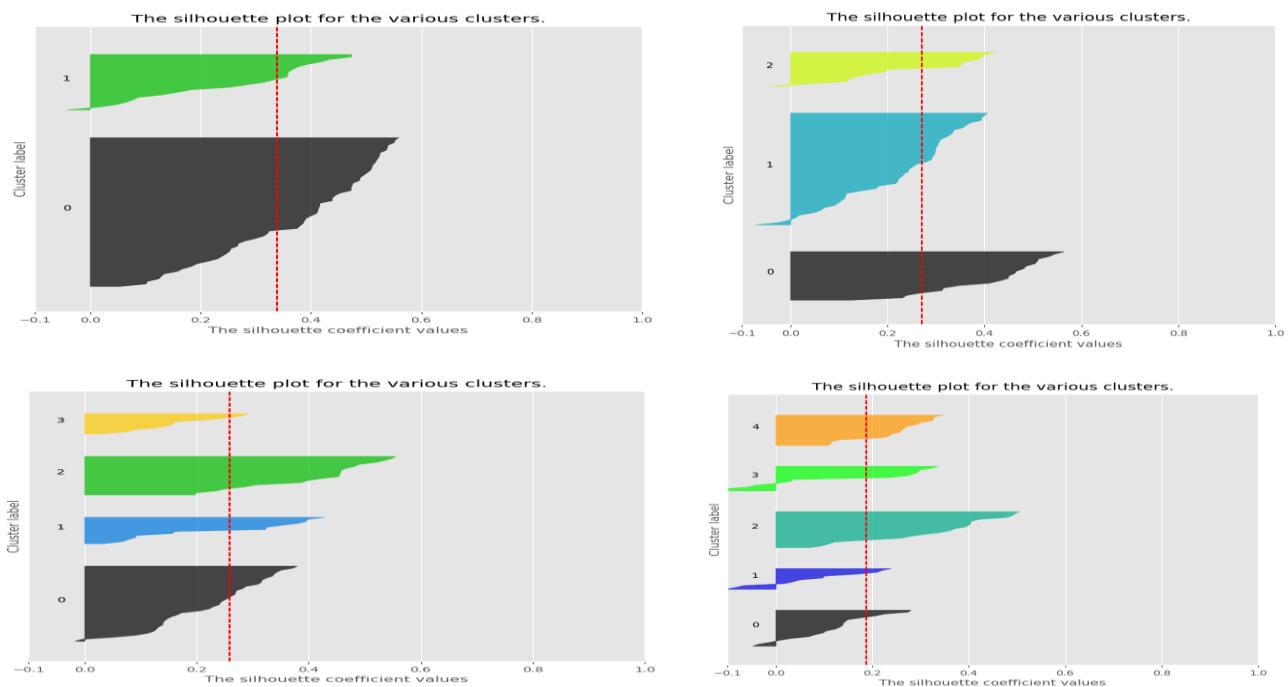
Tras la ejecución del método, se muestra como resultado:



No queda claramente identificado el número óptimo de K , pudiendo oscilar su valor entre 3 y 10 clusters.



Análisis de Silueta



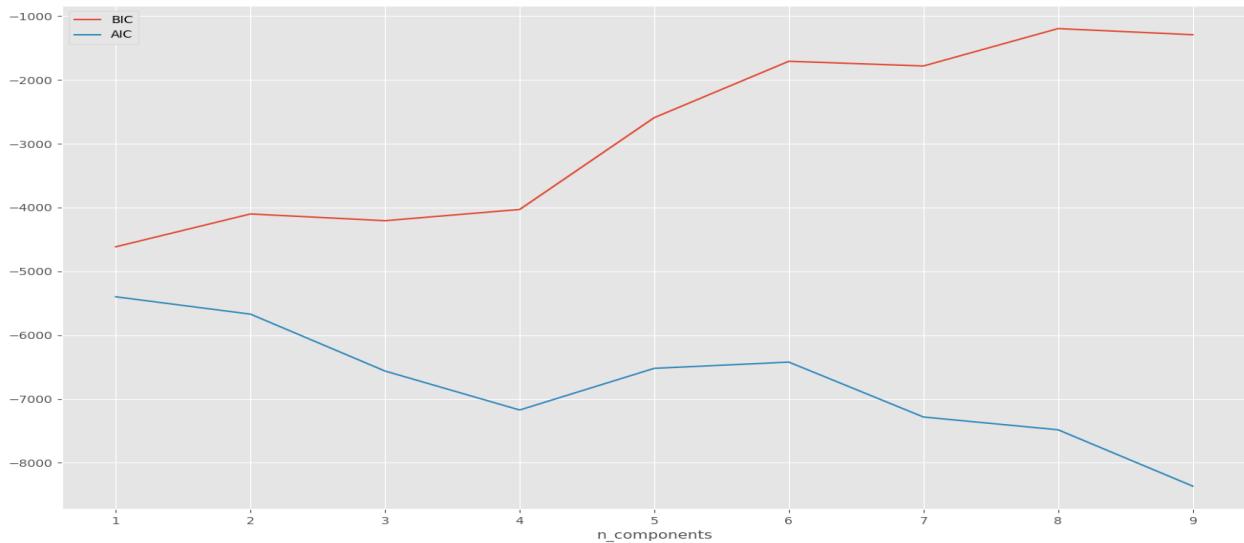
Coeficiente del grado de cohesión de los datos:

Número de Clústers (K)	Coeficiente
2	0. 339304696748
3	0. 271495293762
4	0. 259129942791
5	0. 187196965364

En el resultado gráfico se ve claramente que la mejor agrupación para establecer una K óptima podría ser un valor de $K = 4$.

Gaussian Mixture Models

Se ejecuta el algoritmo con un máximo de 10 componentes, mostrando gráficamente el resultado de las métricas AIC y BIC.



Parece que ambas métricas convergen en un $K = 4$.

Elección de un K Óptimo según las diferentes técnicas aplicadas

En resumen, los resultados obtenidos para cada técnica de cálculo del valor óptimo de K son:

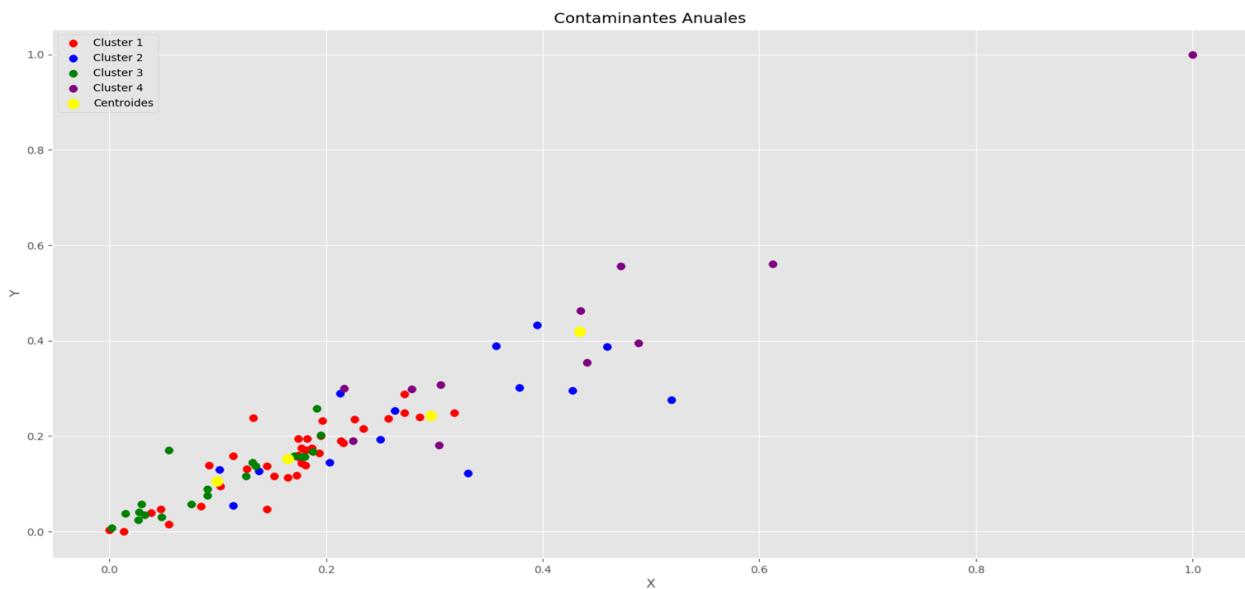
Método	K Óptima
Método Elbow	Oscila entre $K = 3$ y $K = 10$
Análisis de Silueta	$K = 4$
Gaussian Mixture Model	$K = 4$

Si tenemos en cuenta en conjunto todos los resultados, parece que todo converge a un valor óptimo de $K = 4$



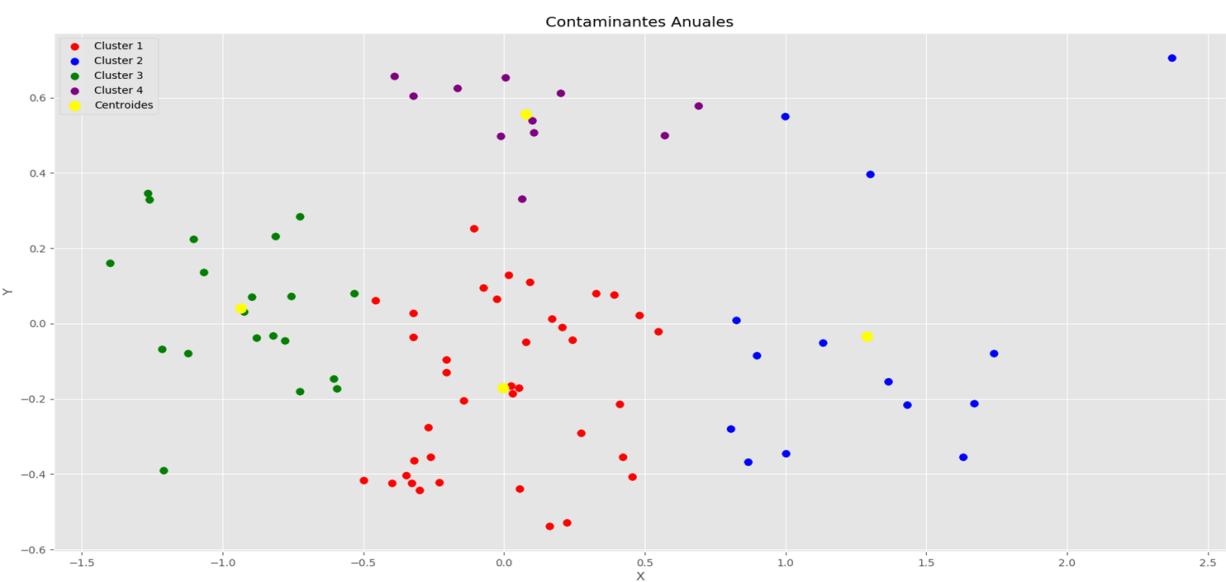
Ejecución de K-Means por Provincia, Municipio, Gases y Trimestre

Una vez calculado el valor de K inicial, procedemos a ejecutar el algoritmo, que como se ha explicado antes, corresponde a un espacio d-dimensional:



Reducción de Dimensionalidad. PCA

Al aplicar *PCA* para reducir las dimensiones del espacio de las muestras obtenemos:



Con este gráfico se ve de forma más clara cómo se distribuyen las muestras entre los diferentes grupos o centroides.

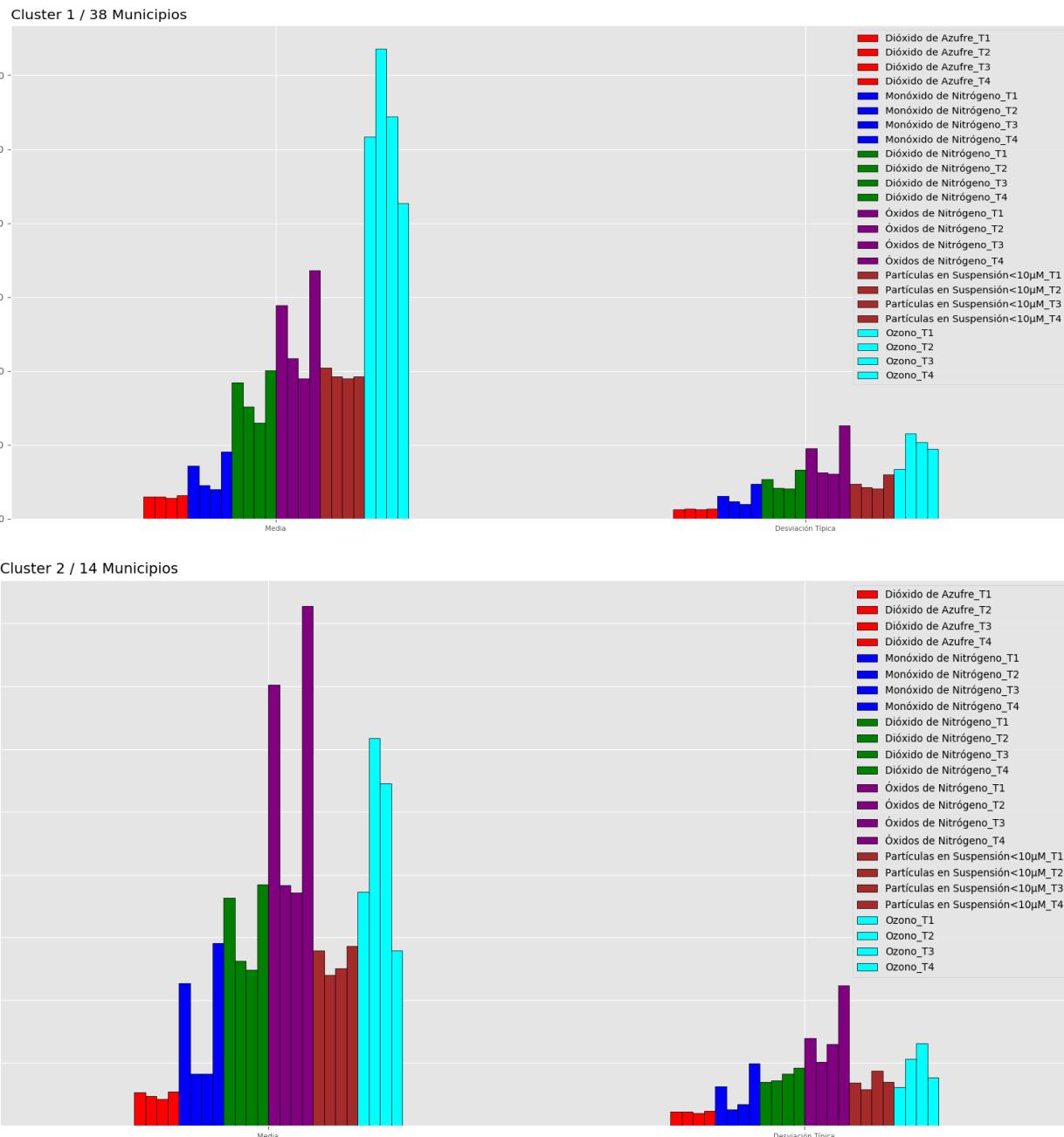


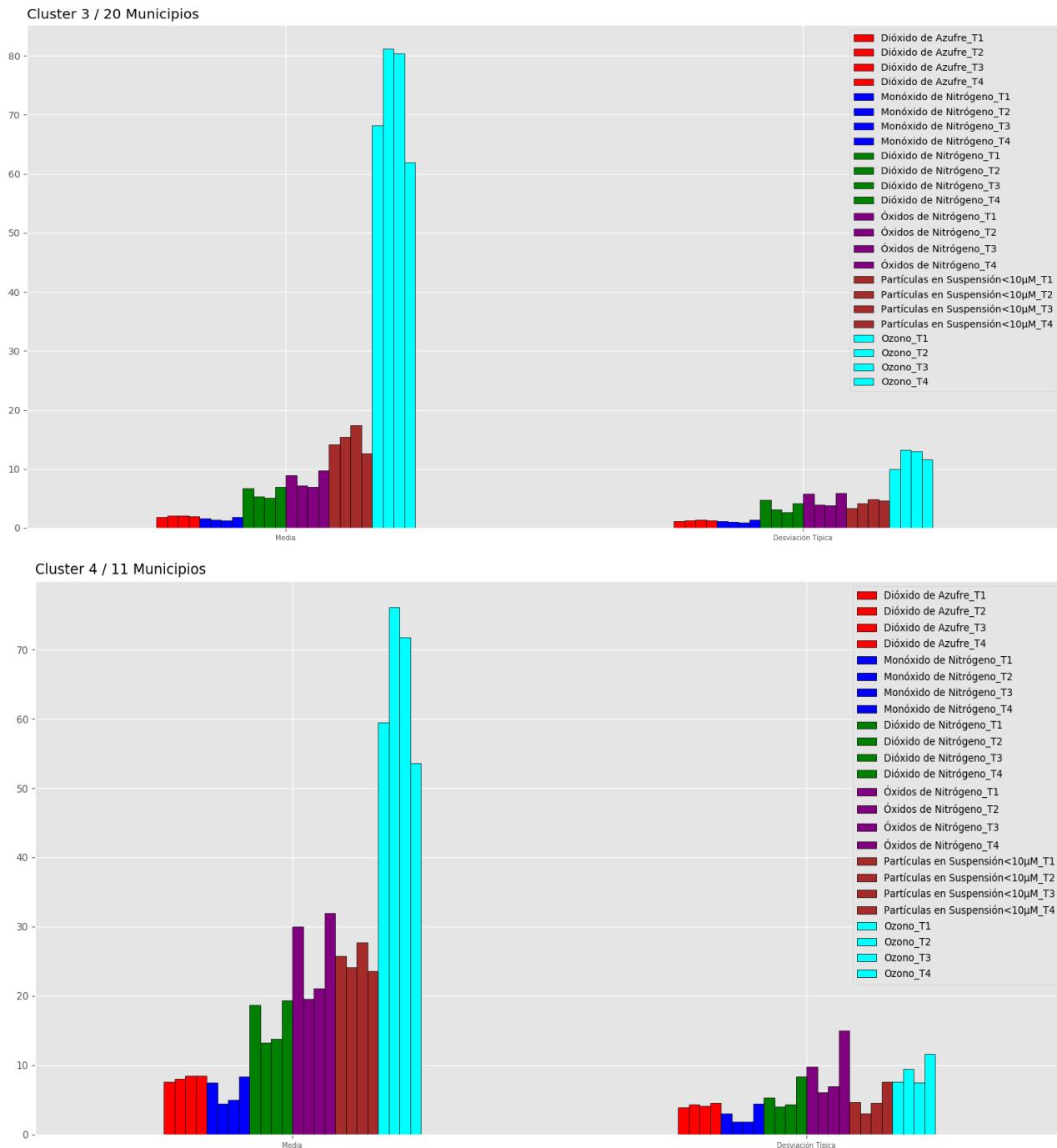
El porcentaje de información explicado por cada uno de los 4 componentes seleccionados es:

61.58 %	10.84 %	7.95 %	7.15 %
---------	---------	--------	--------

Visualización de los Grupos

Una vez ejecutado el método *K-Means*, vamos a visualizar los elementos que componen cada grupo o clúster identificado. Los valores de los contaminantes corresponden a la media y a la desviación típica de los elementos.





Análisis de la Información

Clúster 1

Es el clúster con más municipios, por lo tanto, está ampliamente distribuido geográficamente:

Municipio	Provincia	Municipio	Provincia
Alicante	Alicante	Vigo	Pontevedra



Elche	Alicante	Tenerife	Tenerife
Torrevieja	Alicante	Castro-Urdiales	Cantabria
Mérida	Badajoz	Los Corrales de Buelna	Cantabria
Ibiza	Ibiza	Reinosa	Cantabria
Palma de Mallorca	Palma de Mallorca	Santander	Cantabria
Berga	Barcelona	Torrelavega	Cantabria
Igualada	Barcelona	Segovia	Segovia
Aranda de Duero	Burgos	Vila-seca	Tarragona
Miranda de Ebro	Burgos	Toledo	Toledo
Alcora	Castellón	Alzira	Valencia
Burriana	Castellón	Burjassot	Valencia
Castellón	Castellón	Gandía	Valencia
Onda	Castellón	Paterna	Valencia
A Coruña	A Coruña	Sagunto	Valencia
Santiago de Compostela	A Coruña	Torreblanca	Valencia
Ponferrada	León	Torrente	Valencia
Lugo	Lugo	Media del Campo	Valladolid
Pamplona	Navarra	Alagón	Zaragoza

Se trata de poblaciones de tamaño medio, tanto de costa como de interior, es decir, la localización no influye.

Presenta altos niveles de *Ozono* y moléculas de *Nitrógeno*.

Los valores más altos de *Ozono* en primavera y verano contrastan con los valores más altos de *Nitrógeno* en otoño e invierno y viceversa.

Los valores de *Dióxido de Azufre* y *Material Particulado* son significativos, sin existir valores máximos de ambas sustancias.

Clúster 2

Corresponde principalmente a capitales de provincia, zonas industriales y ubicaciones geográficas que favorecen los vientos y la deposición de material particulado por causa de éstos.

Los municipios son:

Municipio	Provincia	Municipio	Provincia
Barcelona	Barcelona	Móstoles	Madrid
Manlleu	Barcelona	Orense	Orense
El Prat de Llobregat	Barcelona	Avilés	Asturias
San Vicent dels Horts	Barcelona	Gijón	Asturias
Córdoba	Córdoba	Sevilla	Sevilla
Granada	Granada	Valencia	Valencia
Madrid	Madrid	Bilbao	Vizcaya



Presenta los más altos niveles de contaminación de *Nitrógeno y Material Particulado*. Por contra, menor contaminación de *Ozono*.

Esta agrupación es muy similar al *Clúster 4* del anterior estudio realizado para todas las sustancias contaminantes y por lo tanto las posibles causas de estos niveles de contaminación ya han sido expuestas.

Clúster 3

Corresponde a los municipios de mayor contaminación de *Ozono*.

Por contra, menor contaminación de *Nitrógeno, Dióxido de Azufre y Material Particulado*.

Se trata de zonas de montaña y de municipios de tamaño medio y pequeño.

Los municipios son:

Municipio	Provincia	Municipio	Provincia
Alcoy	Alicante	Morella	Castellón
Pinoso	Alicante	San Jorge	Castellón
Badajoz	Badajoz	Víznar	Granada
Zafra	Badajoz	Campisábalos	Guadalajara
Bunyola	Palma de Mallorca	Almonte	Huelva
Mahón	Menorca	Els Torms	Lleida
Cáceres	Cáceres	El Atazar	Madrid
Plasencia	Cáceres	Llanes	Asturias
Toril	Cáceres	San Nicolás del Puerto	Sevilla
Cirat	Castellón	San Pablo de los Montes	Toledo

Clúster 4

Presenta valores significativos de todos los contaminantes.

Junto al *Clúster 2*, tiene los valores más altos de *Material Particulado* y tiene el valor más alto de *Dióxido de Azufre*.

Los municipios son:

Municipio	Provincia	Municipio	Provincia
Albacete	Albacete	Moguer	Huelva
Almería	Almería	Bailén	Jaén
San Fernando	Cádiz	La Robla	León
San Roque	Cádiz	Málaga	Málaga
Puertollano	Ciudad Real	Cartagena	Murcia
Huelva	Huelva		

La alta presencia de *Material Particulado* está influenciada por los vientos



subsaharianos procedentes del norte de África, ya que predomina la zona sur de España.

Además, también aparece el municipio de La Robla, en la zona del Bierzo (León) donde la presencia de contaminantes está originada por la existencia de centrales térmicas y plantas de combustión industrial.

Conclusiones

Coincide como norma general que:

- A mayores niveles de *Ozono*, menores valores de *Dióxido de Azufre, Nitrógeno y Material Particulado*. Estos niveles altos se dan principalmente en primavera y verano.
- Por el contrario, a menores niveles de *Ozono*, mayores valores de los anteriores. Estos máximos niveles se dan en otoño e invierno.

La ubicación geográfica del sur de España coincide con una mayor presencia de *Material Particulado* debido al viento subsahariano.

A continuación, centraremos nuestro estudio únicamente en los contaminantes que son metales

Agrupamiento por Provincia, Municipio, Metales y Trimestre

En este estudio se seguirán los mismos pasos indicados en el punto *Agrupamiento por Provincia, Municipio, Contaminante y Trimestre*.

Utilizando exactamente las mismas técnicas explicadas en el punto anterior, pasaremos a hacer un estudio específico de los siguientes contaminantes:

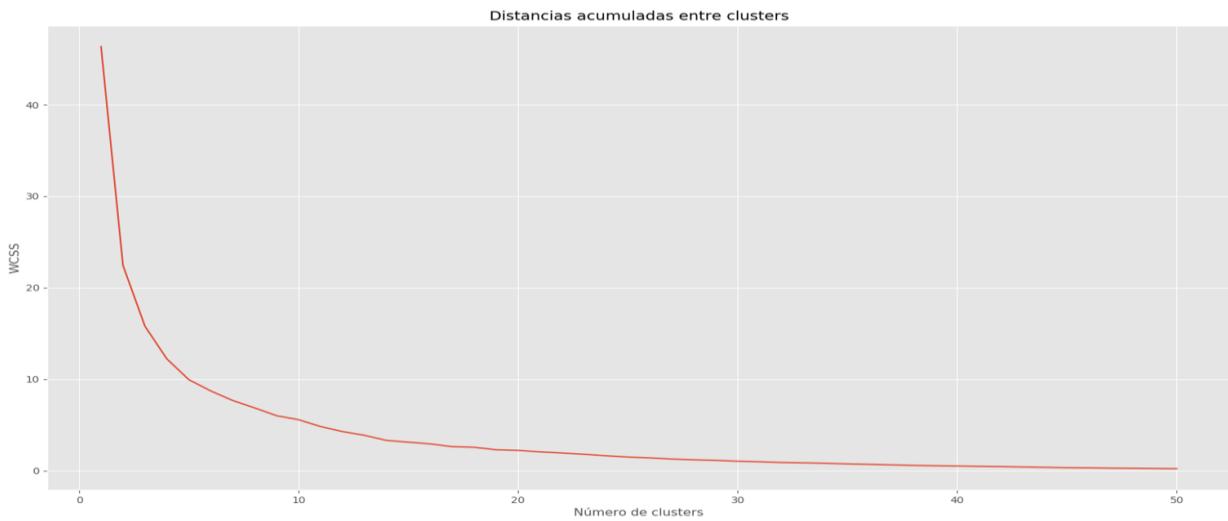
- *Arsénico*
- *Plomo*
- *Cadmio*
- *Níquel*

Al igual que en los estudios anteriores, el primer paso es determinar el valor del parámetro K.

Método Elbow

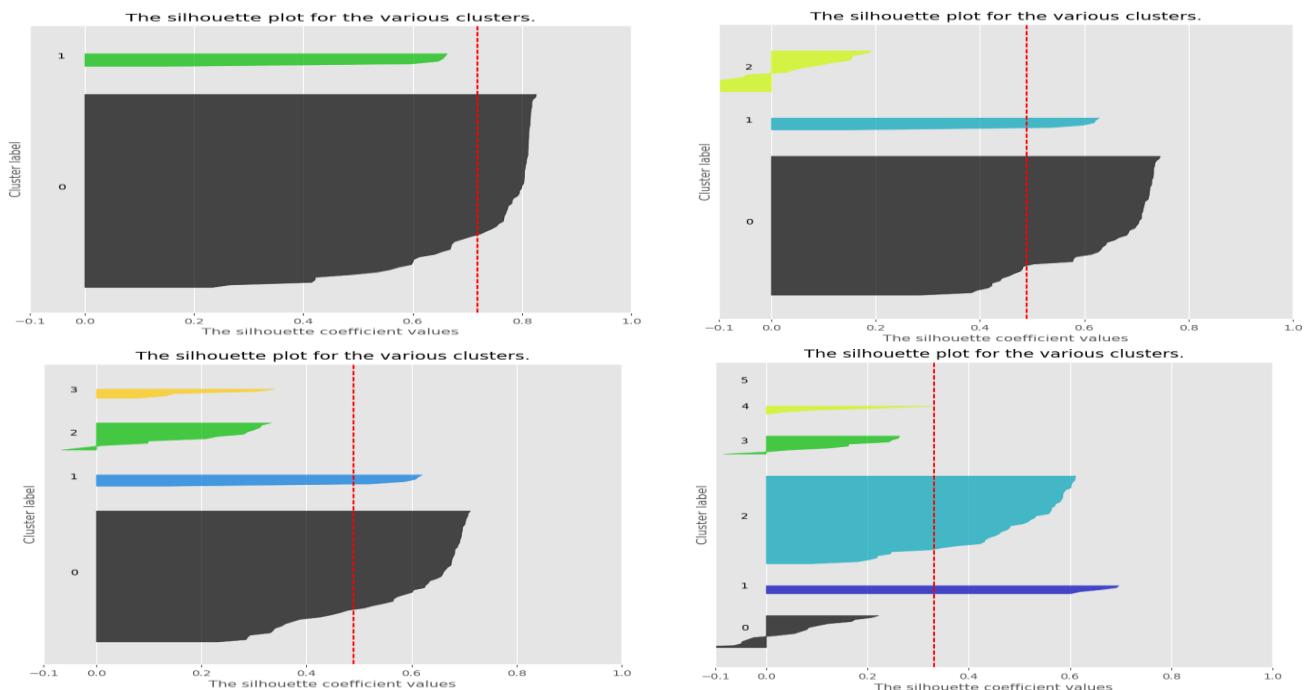
Tras la ejecución del método, se muestra como resultado:





No queda claro el número óptimo de K , pudiendo oscilar su valor entre 3 y 10 *clústers*.

Análisis de Silueta



Coeficiente del grado de cohesión de los datos:

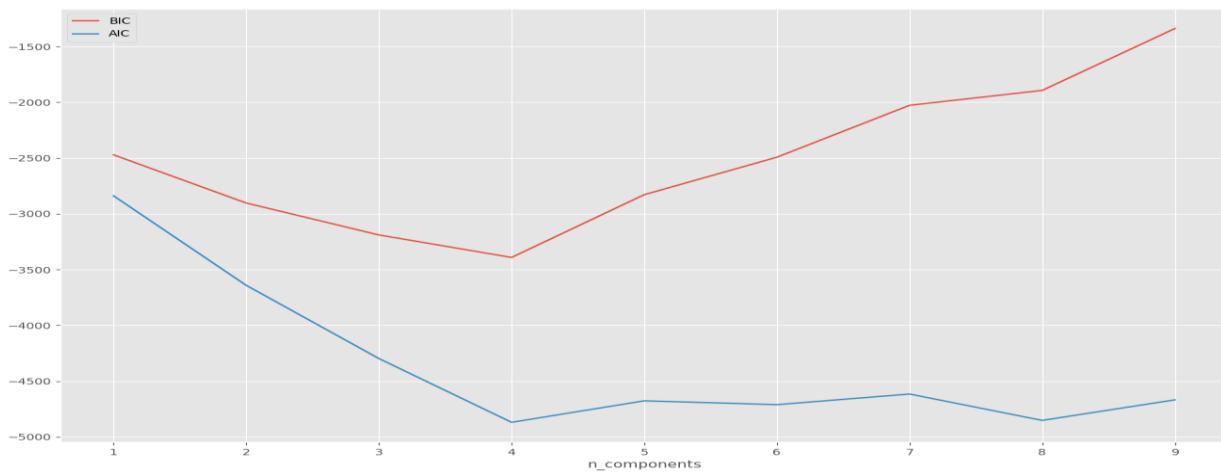
Número de Clústers (K)	Coeficiente
2	0.719173855267
3	0.489325208661
4	0.490123542908
5	0.492420603495
6	0.331876535073



En el resultado gráfico se ve claramente que la mejor agrupación para establecer una K óptima podría ser un valor de $K = 4$.

Gaussian Mixture Models

Se ejecuta el algoritmo con un máximo de 10 componentes, mostrando gráficamente el resultado de las métricas AIC y BIC.



Parece que ambas métricas convergen en un $K = 4$.

Elección de un K Óptimo según las diferentes técnicas aplicadas

En resumen, los resultados obtenidos para cada técnica de cálculo del valor óptimo de K son:

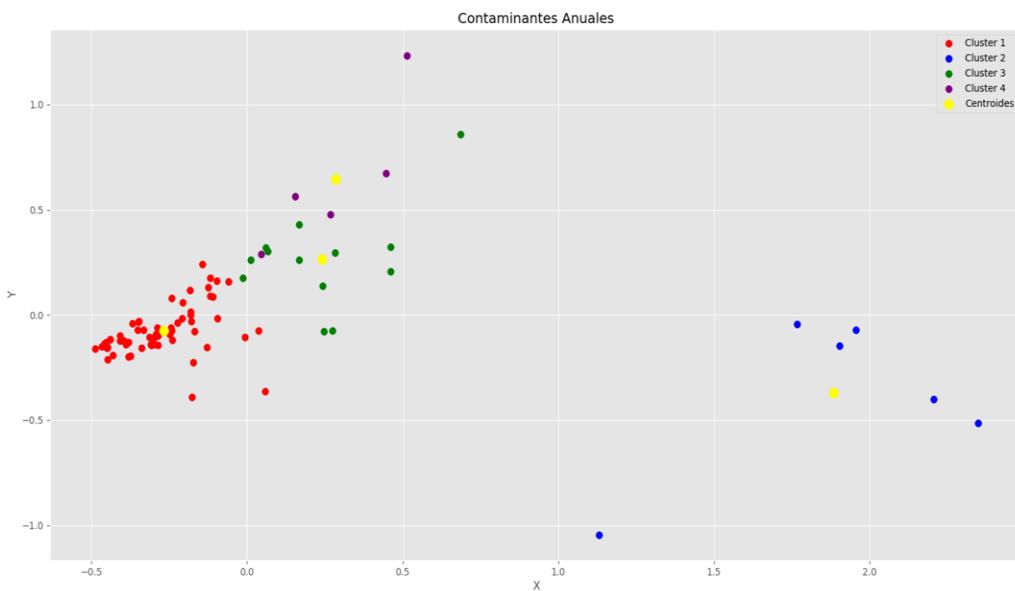
Método	K Óptima
Método Elbow	Oscila entre $K = 3$ y $K = 10$
Análisis de Silueta	$K = 4$
Gaussian Mixture Model	$K = 4$

Si tenemos en cuenta en conjunto todos los resultados, parece que todo converge a un valor óptimo de $K = 4$

Ejecución de K-Means por Provincia, Municipio, Metáles y Trimestre

Una vez calculado el valor de K ejecutamos el algoritmo, mostrando directamente el resultado de aplicar PCA para reducir el espacio de dimensiones de las muestras:





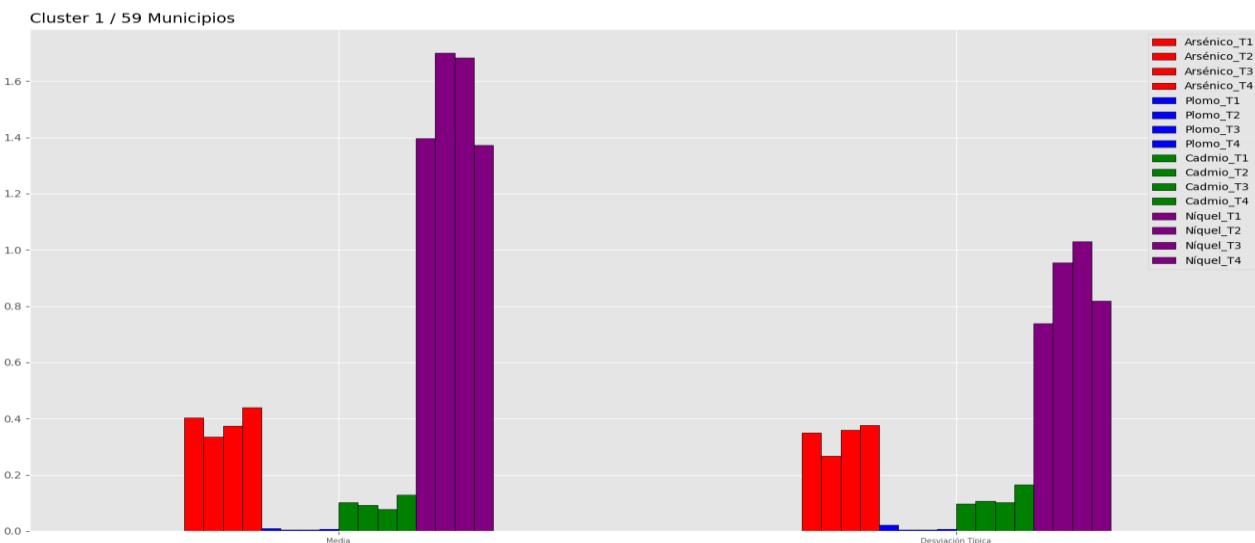
Con este gráfico se ve de forma más clara cómo se distribuyen las muestras entre los diferentes grupos o centroides.

El porcentaje de información explicado por cada uno de los 4 componentes seleccionados es:

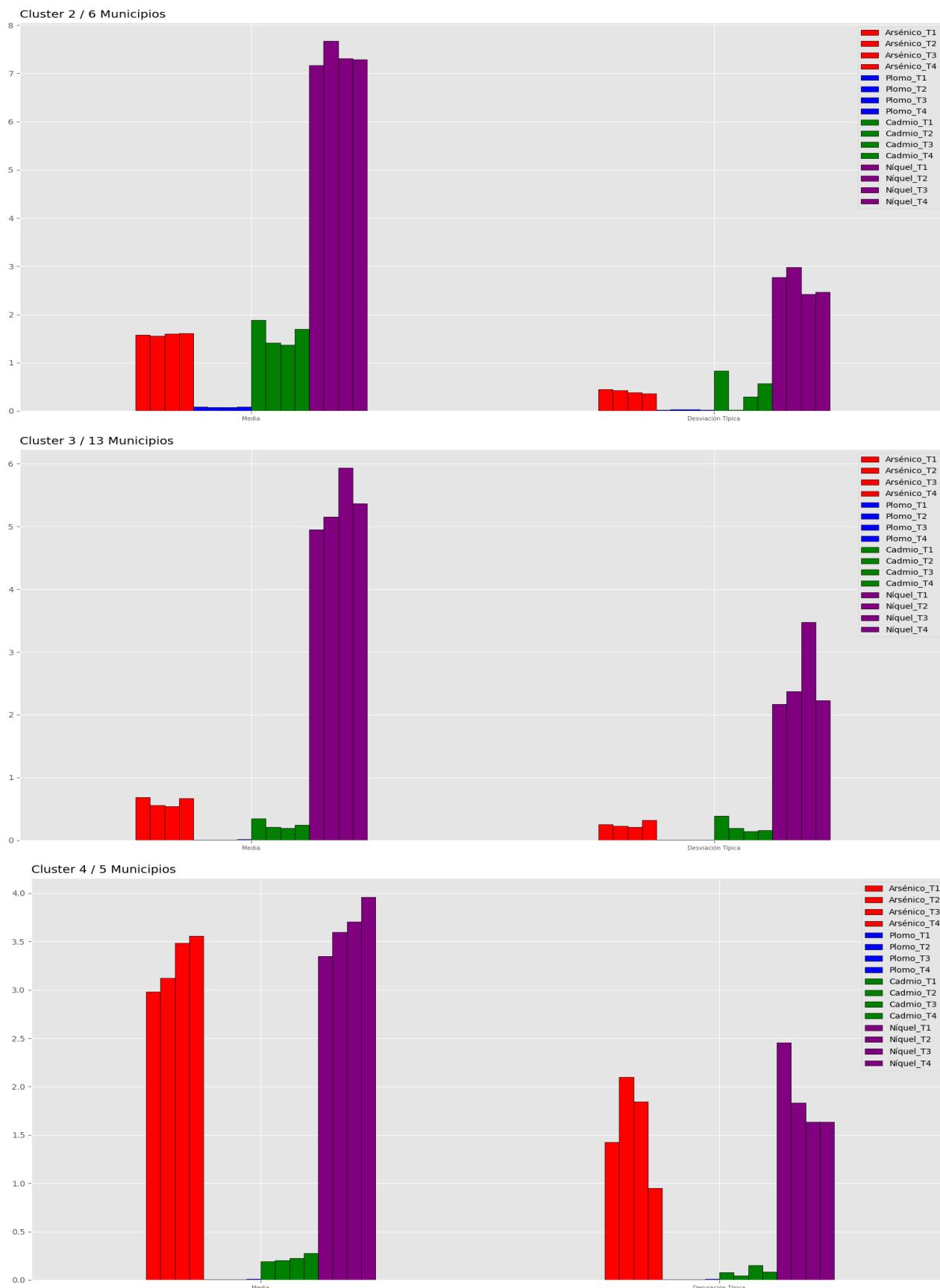
63.28 %	14.82 %	10.42 %	2.89 %
---------	---------	---------	--------

Visualización de los Grupos

Una vez ejecutado el método *K-Means*, vamos a visualizar los elementos que componen cada grupo o clúster identificado. Los valores de los contaminantes corresponden a la media y a la desviación típica de los elementos.



Evaluación sobre la calidad del aire en España



*Análisis de la Información***Clúster 1**

Es el clúster con más municipios de todo el conjunto de datos, por lo tanto, la distribución geográfica es total, no destaca ninguna zona concreta.

Los municipios son:

Municipio	Provincia	Municipio	Provincia
Alicante	Alicante	Madrid	Madrid
Elche	Alicante	Pinoso	Alicante
Torrevieja	Alicante	Castro-Urdiales	Cantabria
Mérida	Badajoz	Los Corrales de Buelma	Cantabria
Alcoy	Alicante	Reinosa	Cantabria
Palma de Mallorca	Palma de Mallorca	Santander	Cantabria
Berga	Barcelona	Torrelavega	Cantabria
Igualada	Barcelona	Segovia	Segovia
Aranda de Duero	Burgos	Badajoz	Badajoz
Miranda de Ebro	Burgos	Toledo	Toledo
Alcora	Castellón	Alzira	Valencia
Burriana	Castellón	Burjassot	Valencia
Castellón	Castellón	Gandía	Valencia
Onda	Castellón	Paterna	Valencia
Móstoles	Madrid	Sagunto	Valencia
Gijón	Asturias	Torrebaja	Valencia
Ponferrada	León	Torrente	Valencia
Sevilla	Sevilla	Medina del Campo	Valladolid
Pamplona	Navarra	Alagón	Zaragoza
Zafra	Badajoz	Bunyola	Palma de Mallorca
Cáceres	Cáceres	Morella	Castellón
Plasencia	Cáceres	San Jorge	Castellón
Toril	Cáceres	Víznar	Granada
Cirat	Castellón	Campisábalos	Guadalajara
Els Torms	Lleida	Almonte	Huelva
El Atazar	Madrid	Albacete	Albacete
Llanes	Asturias	San Fernando	Cádiz
San Nicolás del Puerto	Sevilla	La Robla	León
San Pablo de los Montes	Toledo	Puertollano	Ciudad Real
Cartagena	Murcia		

En general, posee los valores más bajos de todos los contaminantes respecto de los otros clústers.



Los valores de *Plomo* son insignificantes.

Los valores de *Níquel* son más altos en primavera y verano, por el contrario, son más bajos en otoño e invierno.

Los valores de *Arsénico* son a la inversa del Níquel, más altos en otoño e invierno y más bajos en primavera y verano.

Los valores del *Cadmio* son bajos

Clúster 2

Esta agrupación se centra en la comunidad autónoma de Galicia y en Córdoba.

Los municipios son:

Municipio	Provincia
Córdoba	Córdoba
A Coruña	A Coruña
Santiago de Compostela	A Coruña
Lugo	Lugo
Orense	Orense
Vigo	Pontevedra

Se encuentran los valores más altos de *Níquel* y *Cadmio* de todo el conjunto de datos.

La contaminación de *Níquel* en el trimestre 2 (primavera) es mayor al resto que presenta valores más constantes.

El *Cadmio* presenta valores más altos en otoño e invierno.

La presencia de *Plomo* es insignificante.

Los valores de *Arsénico* son significativos, aunque no son los más altos del dataset.

Las causas de esta contaminación en Galicia y Córdoba ya han sido descritas (y son las mismas) en el capítulo *Ejecución de K-Means por Provincia, Municipio, Contaminante y Trimestre* en el apartado *Visualización de los Grupos*, centrado fundamentalmente en la alta presencia de actividad industrial.

Clúster 3

Esta agrupación también presenta amplia distribución geográfica, donde destaca la presencia de ciudades grandes, de costa e industriales.

Los municipios son:

Municipio	Provincia	Municipio	Provincia
Almería	Almería	Málaga	Málaga
Barcelona	Barcelona	Avilés	Asturias
Manlleu	Barcelona	Tenerife	Tenerife
El Prat de Llobregat	Barcelona	Vila-seca	Tarragona



San Roque	Cádiz	Valencia	Valencia
Granada	Granada	Bilbao	Vizcaya
Bailén	Jaén		

La presencia de metales también se ve afectada, por lo tanto, por la existencia de zonas de alta densidad de tráfico rodado y de zonas industriales.

Aunque no son los valores más altos de *Níquel*, sí presenta valores muy elevados y significativos. El *Arsénico* y el *Cadmio* aún en menor medida, son muy bajos. Los valores de *Plomo* son despreciables.

Clúster 4

Esta agrupación está centrada geográficamente en zonas concretas.

Los municipios son:

Municipio	Provincia
Ibiza	Ibiza
Mahón	Menorca
Sant Vicent dels Horts	Barcelona
Huelva	Huelva
Moguer	Huelva

Posee los niveles más altos de *Arsénico*.

La presencia de *Níquel* también es importante, sobre un 50% inferior a los máximos del clúster 2. El *Arsénico* es mayor en la época de verano y otoño, por lo tanto, los más bajos corresponden a invierno y primavera.

El nivel de acumulación de esta sustancia empieza en verano y llega hasta otoño por lo que necesita de una bajada progresiva de las temperaturas para que comience a descender en invierno.

El *Cadmio* tiene poca presencia y el *Plomo* es insignificante.

Conclusiones

Sigue apareciendo la Comunidad Autónoma de Galicia y Córdoba como las más contaminadas en general por sus elevados valores de cualquier tipo de sustancia.

La contaminación de *Níquel* parece que tiene mayor presencia en el periodo de primavera y verano, aunque no es un patrón claro según el clúster 4, circunstancia que también ocurre exactamente igual con el *Arsénico* y el *Cadmio*.

El *Plomo* no tiene presencia, se desconoce si realmente es porque no la tiene o porque no se dispone de datos suficientes de medición de esta sustancia en las estaciones.



6 Análisis Predictivo

En este capítulo se van a detallar los estudios predictivos realizados. Se ha elegido, como objetivo, predecir el valor de la sustancia *Partículas en Suspensión < 10μM (PM₁₀)* al día siguiente (D + 1) a partir de una fecha D de referencia. Los motivos de elegir este contaminante para predecir son fundamentalmente:

- La presencia de este contaminante en el conjunto de datos es constante, para todas las provincias y en todos los períodos del año
- Se trata de un contaminante que por definición puede contener partículas de diferentes sustancias y tipos en diferentes estados, tanto sólido como líquido
- Desde el punto de vista de la salud afecta gravemente y se manifiesta en enfermedades de tipo respiratorio, bronquitis y dolencias de tipo cardiovascular, siendo el precursor de otras enfermedades

El conjunto de datos utilizado para la predicción serán las mediciones de los años 2016, 2017 y 2018.

Una vez decidida la sustancia a predecir, se centrará el estudio en un municipio en concreto

6.1. Elección Municipio para Análisis Predictivo

A continuación, se muestran aquellas poblaciones para elegir una concreta para nuestro estudio, observando las siguientes características:

- Número de mediciones en el conjunto de datos, el total de los 3 años (*)
- Gráfico de las medias de los valores para los meses febrero, abril, agosto y noviembre de ese municipio para los 3 años. Se pretende observar la evolución trimestral del contaminante
- Gráfico de las medias de los valores diarios - semanales (lunes, ..., domingo) de ese municipio para los 3 años

(*) Para los 3 años el número de mediciones totales debe ser $365 \times 3 = 1095 + 1$ (2016 es año bisiesto) = 1096 mediciones

Para cualquier valor inferior faltarán mediciones para algunos días del total de los 3 años.

Municipio	Mediciones
Madrid	1096
Barcelona	1096

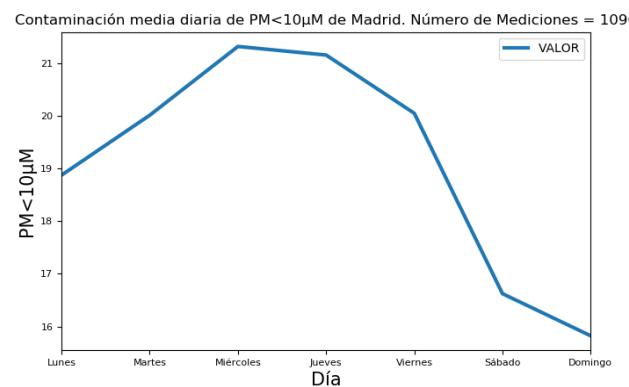
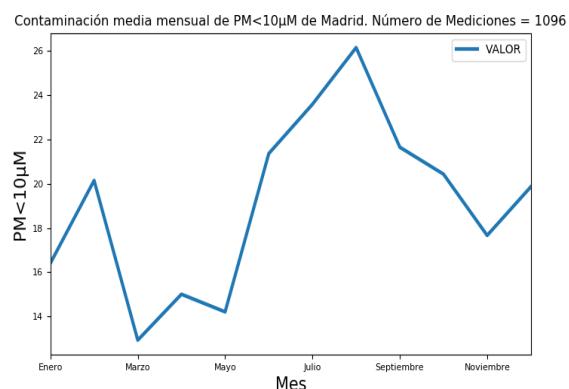


Valencia	1096
Gijón	1096
Avilés (Asturias)	1096
Móstoles (Madrid)	1096
Córdoba	828
Bilbao	1091
Ourense	1090
Sant Vicent dels Horts (Barcelona)	1090

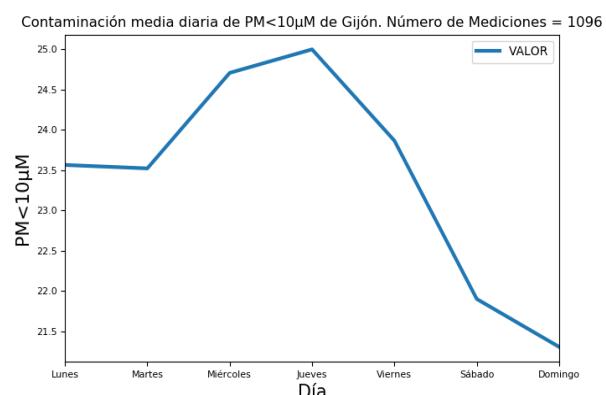
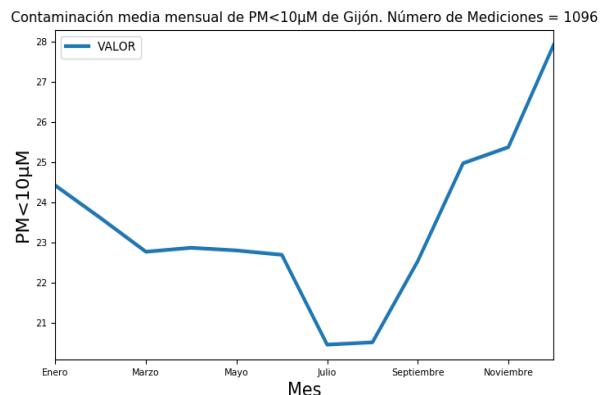
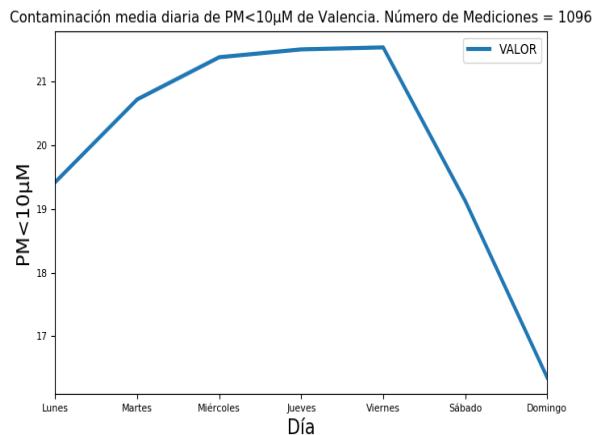
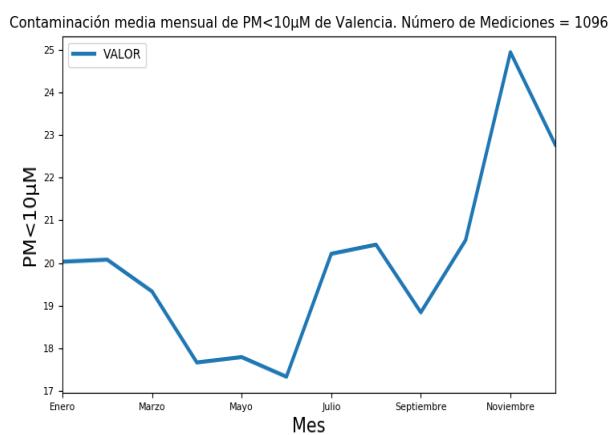
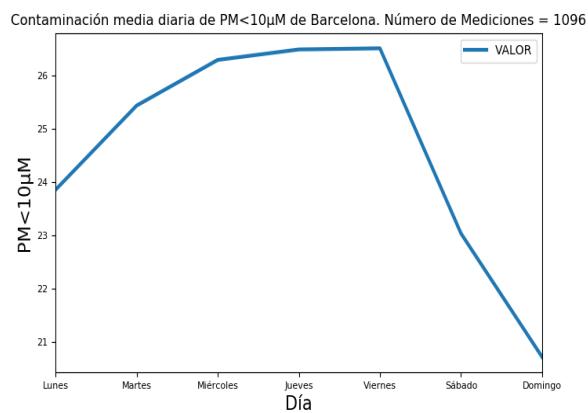
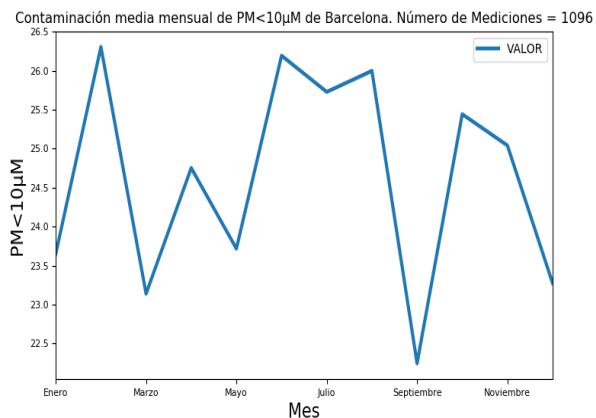
Según estos datos, elegiremos un municipio de la lista con 1096 mediciones, para ello vamos a observar el comportamiento de los contaminantes en los meses indicados y también de forma diaria.

Tenemos el siguiente conjunto de datos:

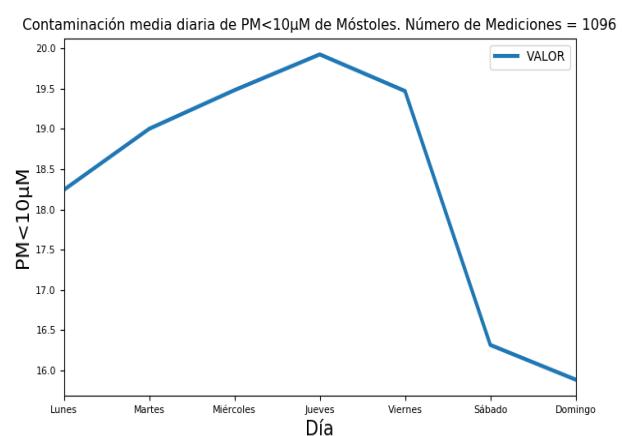
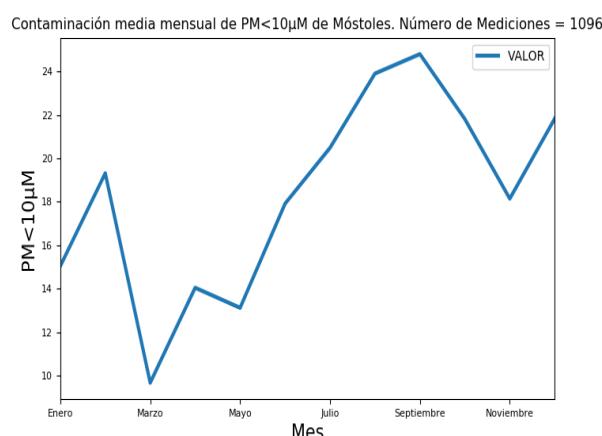
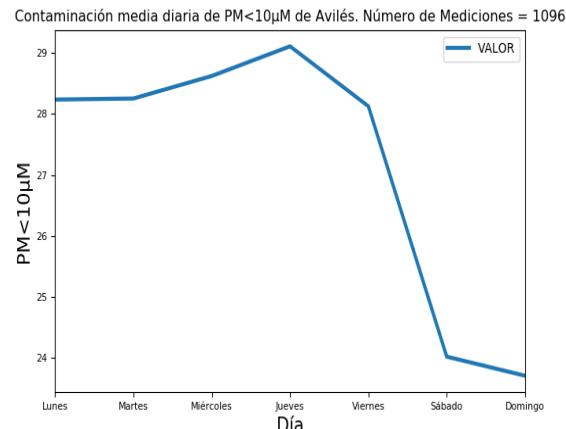
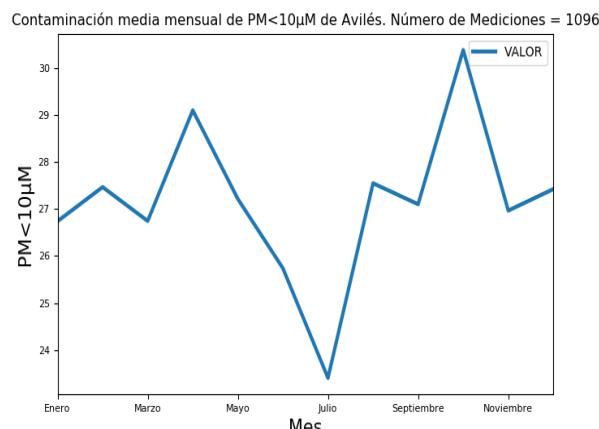
Datos Diarios	Datos Mensuales
Municipio	Municipio
Media Diaria [lunes, ..., domingo] para los años 2016, 2017 y 2018 de Partículas en Suspensión < 10µM	Media Mensual [enero, ..., diciembre] para los años 2016, 2017 y 2018 de Partículas en Suspensión < 10µM



Evaluación sobre la calidad del aire en España



Evaluación sobre la calidad del aire en España



Viendo los gráficos diario y mensual del contaminante en los diferentes municipios, parece interesante estudiar más en detalle las oscilaciones del contaminante en la ciudad de Barcelona, ya que los cambios por trimestre son más pronunciados que en otros municipios.

Por otro lado, los valores diarios del contaminante son generalizados para todos los municipios, mayor contaminación entre semana con una tendencia ascendente hacia el jueves, que empieza a descender de cara al fin de semana.

Se hará una predicción del contaminante Partículas en Suspensión < 10μM sobre la ciudad de Barcelona

Definición de Vista Minable

Antes de comenzar las labores de predicción debemos adaptar la estructura de datos.

Dado que, según los diferentes estudios realizados se ha observado que es importante la época del año y que, en función de ésta, los contaminantes presentan diferentes valores, está claro que la temperatura es un atributo muy importante a tener en cuenta para llevar a cabo una predicción.



Como el estudio es sobre el municipio de Barcelona, necesitamos los datos históricos de las temperaturas medias de los años 2016, 2017 y 2018 de esta ciudad.

Desde la *Agencia Estatal de Meteorología*, existe un apartado *AEMET Open Data*. Se trata de un formulario web para realizar peticiones de datos históricos registrados de toda España por periodos concretos en formatos electrónicos de texto, tales como json, xml, ... etc.

Gracias a esta herramienta, se ha realizado una petición de los datos históricos diarios de la ciudad de Barcelona del periodo comprendido entre el 01/01/2016 y el 31/12/2018, obteniendo el fichero *historico_diario_barcelona.json* con la siguiente estructura:

Dato	Descripción
fecha	Fecha del día en formato AAAA-MM-DD
id	Identificador
nombre	Nombre / Ubicación de la estación
provincia	Provincia de la estación
altitud	Altitud de la estación sobre el nivel del mar
tmed	Temperatura media diaria
prec	Precipitación diaria de 07 a 07
tmin	Temperatura mínima del día
horatmin	Hora y minuto de la temperatura mínima
tmax	Temperatura máxima del día
Horatmax	Hora y minuto de la temperatura máxima
dir	Dirección de la racha de viento máxima
velmedia	Velocidad media del viento
racha	Racha máxima del viento
horaracha	Hora y minuto de la racha máxima
sol	Horas de Insolación
presmax	Presión máxima al nivel de referencia de la estación
horapresmax	Hora de la presión máxima
presmin	Presión mínima al nivel de referencia de la estación
horapresmin	Hora de la presión mínima

Según esta estructura, nos quedamos con los siguientes atributos:

- Fecha
- Temperatura media
- Temperatura mínima
- Temperatura máxima

Fusión datos de contaminante con histórico diario de Barcelona

Estamos en disposición de empezar a dar forma a lo que se conoce como una *vista minable*.

Una *vista minable* no es más que un subconjunto de datos que contiene toda la información necesaria para efectuar un análisis usando técnicas predictivas. Se generarán *vistas minables* definidas a partir del siguiente conjunto de datos que vamos a generar con la siguiente estructura:

Columna	Descripción
TMEDq7	Temperatura media de hace 1 semana
TMED-1	Temperatura media del día D – 1
TMED-2	Temperatura media del día D – 2
TMED-3	Temperatura media del día D – 3
TMED-4	Temperatura media del día D – 4
TMED-5	Temperatura media del día D – 5
TMED-6	Temperatura media del día D – 6
TMED-7	Temperatura media del día D – 7
TMIN	Temperatura mínima del día D
TMAX	Temperatura máxima del día D
PMq7	Valor medio de la sustancia contaminante hace 1 semana
PM-1	Valor de la sustancia contaminante del día D – 1
PM-2	Valor de la sustancia contaminante del día D – 2
PM-3	Valor de la sustancia contaminante del día D – 3
PM-4	Valor de la sustancia contaminante del día D – 4
PM-5	Valor de la sustancia contaminante del día D – 5
PM-6	Valor de la sustancia contaminante del día D – 6



PM-7	Valor de la sustancia contaminante del día D – 7
D (1)	Día a predecir en formato DD/MM/YYYY
PM	Valor a predecir de la sustancia contaminante del día D
FESTIVO (2)	Indicador de si el día D es festivo o no (1/0)

Para integrar este conjunto de datos, debemos, como paso inicial, fusionar los datos del contaminante *Partículas en Suspensión < 10μM* con los datos históricos obtenidos de *Aemet* de la ciudad de Barcelona en función de la fecha, del día "D".

Fecha	Partículas en Suspensión < 10μM	Fecha	T. Mínima	T. Máxima	T. Media
DD/MM/YYYY	X1.0X	DD/MM/YYYY	T1.0X	T2.0X	T3.0X
...
DD/MM/YYYY	Xn.0X	DD/MM/YYYY	T1n.0X	T2n.0X	T3n.0X



Fecha	Partículas en Suspensión < 10μM	T. Mínima	T. Máxima	T. Media
DD/MM/YYYY	X1.0X	T1.0X	T2.0X	T3.0X
...
DD/MM/YYYY	Xn.0X	T1n.0X	T2n.0X	T3n.0X

A partir de esta definición inicial de la *vista minable* podemos ir calculando el resto de los atributos TMED-n y PM-n en función de la fecha "D" para tener la *vista minable* completa.

(1) Tomando como referencia esta fecha "D", la predicción se realizará para una fecha D + 1. El sentido de incluir este campo como parte de la *vista minable* es únicamente a nivel informativo, ya que sólo se utilizará para cruzar los datos obtenidos de Aemet con nuestro conjunto de datos del contaminante *Partículas en Suspensión < 10μM*, a nivel de ejecución de los algoritmos predictivos no formará parte de la *vista minable*.

(2) Ha sido necesario obtener el calendario de festivos de la ciudad de Barcelona de los 3 años, festividades nacionales, autonómicas y locales. Según los resultados de los estudios anteriores, los períodos de fin de semana y/o festivos influyen claramente en el valor de los contaminantes, por lo tanto, es interesante incluir este atributo en la *vista minable*.

El rango de fechas de nuestra *vista minable* irá desde el 11/01/2016 hasta el 31/12/2018, ya que el atributo Temperatura Media de la semana anterior del día D (TMEDq7) necesita obtener datos entre los 7 y 14 días anteriores al día "D".

Disposición de los datos para la predicción

De la misma forma que se han aplicado en el capítulo del *Análisis Descriptivo* unos pasos previos a las técnicas de clasificación con *K-Means*, también es necesario hacerlo para las técnicas de predicción:



- Eliminar Datos Faltantes. Eliminar aquellas filas de la *vista minable* que contengan algún valor nulo
- Dado que la *vista minable* está compuesta por valores de temperatura y sustancia contaminante, los atributos ya son numéricos
- Normalizar datos. Como esta técnica permite ajustar valores numéricos de diferentes escalas, es conveniente aplicar esta transformación en los datos, además de que algunos algoritmos de aprendizaje funcionan mejor con atributos numéricos normalizados entre 0 y 1, como es el caso de la *regresión lineal* y las *redes neuronales*.

Tal y como se ha descrito en el capítulo 2.3 *Técnicas de aprendizaje supervisado*, habitualmente se utilizan datos históricos para crear un modelo matemático que capture patrones y tendencias, para, aplicando ese modelo a partir de datos reales, poder llevar a cabo una predicción con garantías.

En nuestro estudio tenemos un único conjunto de datos, la *vista minable*, y debemos dividirlo en un conjunto para generar los modelos y en otro conjunto para predecir a partir de datos reales, es lo que se denomina datos de *training* y datos de *test* respectivamente. Nuestra división será la siguiente.

Datos de Training

Serán todas aquellas filas de la *vista minable* con fecha menor o igual al 30 de junio del 2018, es decir, periodo 2016 y 2017 completo y 6 meses de 2018. Este conjunto de datos se utilizará para entrenar nuestros modelos matemáticos de tal forma que deberán ser capaces de establecer patrones de los posibles valores a predecir de las *Partículas en Suspensión*.

Datos de Test

Serán todas aquellas filas de la *vista minable* con fecha mayor o igual al 1 de julio de 2018, es decir los 6 meses restantes del 2018.

Este conjunto de datos se utilizará para verificar si la predicción de nuestros modelos matemáticos se ajusta a estos valores reales.

Una vez la vista minable está formada y los datos de training y test subdivididos es momento de aplicar las técnicas de predicción

Técnicas de Predicción aplicadas

Para este estudio se va a predecir el valor del contaminante de *Partículas en Suspensión*, un valor numérico real, es lo que se conoce como un problema de regresión, tal y como se ha explicado en el capítulo 2.



Las técnicas predictivas que se van a utilizar son:

- Regresión Lineal
- Árbol de Decisión
- Random Forest
- Red Neuronal

Para poder comparar los resultados de cada algoritmo se ha trabajado con diferentes modelos de la *vista minable*, de esta forma se puede observar cómo de importantes son los atributos elegidos y si aportan información valiosa o no.

Para evaluar los diferentes modelos sobre el conjunto de test usaremos las siguientes medidas de evaluación aplicables a problemas de regresión:

- MSE. Error cuadrático medio. Es un valor numérico estimador que determina cómo de buena es la técnica predictiva aplicada, calculando el promedio de error al cuadrado entre el valor predicho y el valor real. Es un porcentaje y cuanto más bajo es este valor mejor es la predicción realizada ya que el error es menor.
- MAE. Error absoluto medio. El concepto es similar al estimador anterior. Calcula, en valor absoluto, la diferencia entre el valor predicho y el valor real. Al igual que el anterior, los valores más bajos corresponden a mejores predicciones.

A continuación, se presentan los resultados de los modelos predictivos por diferentes estructuras de la vista minable.

Modelo de referencia. Baseline

Una forma de corroborar que las técnicas a aplicar son correctas, es que han de mejorar siempre un *modelo de referencia base*. El *modelo de referencia base* elegido es la media de los *datos de training* frente a los *datos reales de test*.

Baseline	
MSE	42.38673
MAE	5.21915

Este resultado se comparará con los resultados que se obtengan utilizando la *vista minable*.



Vista 1. Valores medios para los 7 días, sin tener en cuenta indicador de festivo

ALGORITMO	MSE	MAE	Vista Minable
Regresión Lineal	0.00256	0.03728	TMEDq7
Árbol de Decisión	0.00391	0.04864	TMED-n (n = 1 ... 7)
Random Forest	0.00322	0.04287	TMIN
Red Neuronal	0.00111	0,02248	TMAX PMq7 PM-n (n = 1 ... 7) PM

Vista 2. Se excluyen valores medios de la semana anterior (q7) e indicador de festivo

ALGORITMO	MSE	MAE	Vista Minable
Regresión Lineal	0.00259	0.03763	TMED-n (n = 1 ... 7)
Árbol de Decisión	0.00391	0.04864	TMIN
Random Forest	0.00321	0.04275	TMAX PM-n (n = 1 ... 7)
Red Neuronal	0.00080	0,02055	PM

Vista 3. Todos los atributos, para solamente 3 días, sin tener en cuenta indicador de festivo

ALGORITMO	MSE	MAE	Vista Minable
Regresión Lineal	0.00251	0.03700	TMEDq7
Árbol de Decisión	0.00395	0.04802	TMED-n (n = 1 ... 3)
Random Forest	0.00338	0.04399	TMIN
Red Neuronal	0.00133	0,02708	TMAX PMq7 PM-n (n = 1 ... 3) PM

Vista 4. Todos los valores medios para los 7 días, incluyendo indicador de festivo

ALGORITMO	MSE	MAE	Vista Minable
Regresión Lineal	0.00219	0.03473	TMEDq7
Árbol de Decisión	0.00389	0.04810	TMED-n (n = 1 ... 7)
Random Forest	0.00292	0.04118	TMIN
Red Neuronal	0.00084	0.02192	TMAX PMq7 PM-n (n = 1 ... 7) PM Festivo



Vista 5. Sin valores medios semana anterior (q7) y valores medios para 3 días con indicador de festivo

ALGORITMO	MSE	MAE	Vista Minable
Regresión Lineal	0.00222	0.03530	TMED-n (n = 1 ... 3)
Árbol de Decisión	0.00387	0.04800	TMIN TMAX
Random Forest	0.00295	0.04137	PM-n (n = 1 ... 3)
Red Neuronal	0.00099	0.02276	PM Festivo

Vista 6. Todos los valores medios para 3 días con indicador de festivo

ALGORITMO	MSE	MAE	Vista Minable
Regresión Lineal	0.00220	0.03459	TMEDq7
Árbol de Decisión	0.00381	0.04825	TMED-n (n = 1 ... 3) TMIN TMAX
Random Forest	0.00320	0.04344	PMq7 PM-n (n = 1 ... 3) PM Festivo
Red Neuronal	0.00069	0.01881	

Modelo de Referencia Baseline vs Modelos de Vista Minable

Como se puede observar, los resultados del modelo baseline son peores que nuestra predicción y es lógico. En este modelo base no se tienen en cuenta factores como las temperaturas y los valores medios de los contaminantes de los días anteriores, mínimas, máximas e información de días festivos, no existe ningún tipo de precisión.

Resumen Modelos de Vista Minable

Como se puede ver, en general para todos los métodos predictivos utilizados, los resultados son muy buenos, del orden de error de 0.2 - 0.3 %

- Los métodos *regresión lineal* y *random forest* donde mejor funcionan es en la *Vista Minable*. Esta vista contiene el máximo de información posible de la estructura de datos, ya que incluye todos los valores medios para los 7 días, temperatura y contaminante medio de la semana anterior e indicador de día festivo.

ALGORITMO	MSE	%
Regresión Lineal	0.00219	0.219 %
Random Forest	0.00292	0.292 %



- Los métodos *árbol de decisión* y *red neuronal* donde mejor funcionan es en la *Vista Minable 6*. Esta vista difiere de la anterior en que coge solamente las temperaturas medias de los 3 días anteriores. El resto es igual, temperatura y contaminante medio de la semana anterior e indicador de día festivo.

ALGORITMO	MSE	%
Árbol de Decisión	0.00381	0.381 %
Red Neuronal	0.00069	0.069 %

Como ha quedado demostrado, no siempre añadir más atributos es mejor según qué método de predicción se utilice:

- Se aprecia una mejora con la presencia del indicador de día festivo
- Según qué técnica, funciona mejor con temperaturas medias de 3 días que con 7 días hacia atrás

Este estudio es con un volumen de datos muy limitado, por lo tanto, computacionalmente no presenta ningún problema, siendo la *red neuronal* la técnica que mejores resultados ofrece, sin embargo, para problemas más complejos con una elevada carga de datos, esta técnica puede presentar problemas de complejidad de implementación, coste y computación.

Por otro lado, la *regresión lineal* computacionalmente es más estable y eficiente, tiene menor coste y debe ser valorado como una muy buena alternativa de estudio.

A partir de aquí, debe ser criterio del analista que estructura de datos montar y qué técnica predictiva aplicar, aunque lo ideal siempre es poder comparar y contrastar diferentes técnicas y modelos de datos.

7 CONCLUSIÓN

Según las diferentes técnicas de clasificación y predicción aplicadas, los resultados obtenidos exponen de forma clara diferentes aspectos:

El tamaño de la población influye en los valores de las sustancias contaminantes, siendo los núcleos urbanos más poblados los que mayor contaminación presentan. La contaminación evoluciona de forma descendente en zonas interurbanas y rurales (excepto el *Ozono* que es a la inversa).

El estudio inicial de datos horarios y diarios indica de forma clara la alta presencia de contaminación en zonas de alta densidad de actividad industrial y de tráfico, así como gran presencia de material particulado ampliamente distribuido geográficamente por toda España, incluidas las islas. El estudio de datos unificado corrobora esta circunstancia.



La contaminación por tráfico queda ligada de forma directa a los valores de las diferentes moléculas de *Nitrógeno*, y de manera indirecta por la presencia de *Material Particulado*, *Arsénico*, *Dióxido de Azufre* y otros contaminantes, sustancias que también intervienen en el proceso de combustión de los vehículos a motor.

La contaminación por Ozono se produce, en gran medida, como consecuencia de la reacción química de las diferentes moléculas de *Nitrógeno*, *el Monóxido de Carbono*, *el Metano* y otros compuestos volátiles dando lugar a este contaminante.

Los extrarradios de las ciudades y las áreas rurales registran concentraciones más elevadas que los centros urbanos debido al transporte atmosférico que experimentan las emisiones urbanas.

La formación del *Ozono* troposférico no es inmediata, mientras se producen las diferentes reacciones químicas, los precursores que dan lugar al *Ozono* pueden ser arrastrados a varios kilómetros de distancia. Sin embargo, en las ciudades se produce de forma inversa, haciendo que el *Ozono* se degrade al reaccionar con el *Nitrógeno* y, por lo tanto, permitiendo que en los núcleos urbanos el *Ozono* se mantenga en equilibrio con el *Nitrógeno*.

Esto explica porque la relación de los niveles de *Nitrógeno* y *Ozono* son a la inversa, cuando uno presenta valores altos, el otro tiene valores más bajos y viceversa.

La contaminación por fuente de emisión de tipo industrial está localizada geográficamente en zonas con alta densidad de industria donde no se ha aplicado correctamente un plan urbanístico acorde con un desarrollo de esta actividad, destacando industria relacionada con la producción y procesado de materiales, sustancias químicas y centrales térmicas.

Destaca muy por encima del resto la Comunidad Autónoma de Galicia.

De forma bastante generalizada, los trimestres de mayor contaminación de *Nitrógeno* son los trimestres 1 y 4, periodo que coincide con otoño e invierno, es decir, mayor desplazamiento de vehículos, actividad industrial y consumo energético.

La cercanía del sur de España con África indudablemente también influye en la presencia de *Material Particulado* impulsado por los vientos subsaharianos.

En cuanto a la contaminación de los diferentes metales, los valores del Plomo no son significativos en el estudio.

Sí son significativos los valores del resto de sustancias: *Arsénico*, *Cadmio* y *Níquel*.

Destacan los valores de la Comunidad Autónoma de Galicia. Sin duda, como se ha explicado antes, el tráfico y la actividad industrial influyen, no sólo en la emisión de gases, sino también en la de metales, pudiéndose generalizar en otras provincias o municipios con valores altos de estos contaminantes.

Desde el punto de vista descriptivo, se podría incluir un estudio adicional que determinara si algún municipio supera los umbrales o valores límite permitidos por la legislación vigente para algún contaminante, tal y como indica el Anexo II de este documento.

Desde el punto de vista predictivo, los valores estimadores obtenidos con tanta precisión obedecen a la aplicación de los modelos matemáticos con una *vista minable*. Esta técnica, sin duda, permite



acerarse en gran medida a una realidad, los valores de temperatura y contaminación en períodos anteriores influyen en los valores de contaminación del día de mañana.

Conociendo más profundamente el modelo de negocio, seguro que es posible tener en cuenta otras variables no presentes en este estudio tales como la presión atmosférica, velocidad del viento, datos de precipitación, humedad y un largo etc., que podrían mejorar los resultados obtenidos, aunque según se menciona en el apartado de la predicción, no siempre añadir más información va a mejorar los resultados, en este sentido siempre habrá que aplicar diferentes técnicas y modelos para comparar resultados.

En relación a la calidad de los datos se han visto varios aspectos que merece la pena destacar:

- La información, en general, es escasa, teniendo en cuenta la gran cantidad de estaciones de medición existentes por todo el país, ha quedado de manifiesto que no todas las estaciones registran información de todos los contaminantes, aspecto que hace insuficiente poder llevar a cabo un estudio con resultados que ofrezcan verdaderas garantías
- Asimismo, existen, de forma muy generalizada, muy pocos datos de contaminantes tales como el *Monóxido de Carbono*, *Partículas en Suspensión <2.5 μM*, *Benzoapireno* y el *Benceno*, sustancias excluidas del *análisis descriptivo* que son contaminantes peligrosos y fundamentales considerados de efecto invernadero
- También indicar la alta presencia de valores nulos en diferentes frecuencias de medición de las estaciones, datos diarios y horarios, para todos los municipios, durante todo el año. Este hecho también aumenta la complejidad de poder hacer un estudio fiable
- La mejora de estos aspectos permitiría, sin duda, la elaboración de un estudio más completo y adecuado a nuestra realidad

Tal y como se ha descrito en el trabajo realizado, queda claramente demostrado que es, sin duda, la actividad humana la que está contribuyendo al cambio climático.

Estudios científicos sugieren que la contaminación está asociada con incrementos en la morbi-mortalidad de la población, al creciente desarrollo de enfermedades respiratorias y a la aparición de alergias entre la población infantil.

El pequeño tamaño de este material particulado y la presencia de gases y metales tóxicos en el aire que respiramos hace posible su irrupción en nuestro aparato respiratorio, llegando a nuestro torrente sanguíneo. Asimismo, su ligereza y volatilidad hace que permanezcan por más tiempo en el aire y facilita su transporte por el viento a grandes distancias.

También es el efecto que tiene en el medio ambiente, como la contaminación de las aguas, el subsuelo, las plantas, la agricultura y la ganadería, poniendo de manifiesto que estas sustancias entran en nuestra cadena alimenticia influyendo en nuestra salud.



La aportación de la propia naturaleza es, en proporción, mucho menor que la acción del ser humano.

8 Bibliografía

- [1] *Datos abiertos proporcionados por el Ministerio para la Transición Ecológica.*
<https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/>
- [2] *Causas del Cambio Climático y del Calentamiento Global.* Oxfam Intermón.
<https://blog.oxfamintermon.org/causas-del-cambio-climatico-calentamiento-global>
- [3] *Metodología CRISP-DM* Wirth, R & Hipp, J. (2000, April): Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). London, UK: Springer-Verlag.
- [4] *Introducción al Business Intelligence.* Fernando Martínez Plumed
- [5] *Introducción a la Minería de Datos.* Fernando Martínez Plumed. Jose Hernández Orallo. Departamento de Sistemas Informáticos y Computación. Escuela Superior de Ingeniería Informática. Universidad Politécnica de Valencia
- [6] *El Proceso de Extracción del Conocimiento.* Fernando Martínez Plumed. Jose Hernández Orallo. Departamento de Sistemas Informáticos y Computación. Escuela Superior de Ingeniería Informática. Universidad Politécnica de Valencia
- [7] *Introducción al Aprendizaje Automático I, II y III.* Jon Ander Gómez. Pattern Recognition and Human Language Technologies Research Center. Departamento de Sistemas Informáticos y Computación. Escuela Técnica Superior de Ingeniería Informática. Universidad Politécnica de Valencia
- [8] *Herramientas Estadísticas para Big Data. Introducción a la Inferencia Estadística Muestreo y Preproceso de datos.* Elena Vázquez. Departamento de Estadística e Investigación Operativa Aplicadas y Calidad. Universidad Politécnica de Valencia
- [9] *Herramientas Estadísticas para Big Data. Evaluación de modelos e implantación.* Elena Vázquez Barrachina, Mónica Clemente Císcar. Ana Debón Aucejo. Departamento de Estadística e Investigación Operativa Aplicadas y Calidad. Universidad Politécnica de Valencia
- [10] *Herramientas Estadísticas para Big Data. Clasificadores.* Elena Vázquez Barrachina, Mónica Clemente Císcar. Ana Debón Aucejo. Departamento de Estadística e Investigación Operativa Aplicadas y Calidad. Universidad Politécnica de Valencia



[11] *Evaluación de la Calidad del Aire. Umbrales y valores límite.*
<https://www.boe.es/boe/dias/2002/10/30/pdfs/A38020-38033.pdf>

[12] *Estadística y Machine Learning con R. Francisco Parra.*
<https://bookdown.org/content/2274/portada.html>

[13] *Mapas con R. Web de Franz Jimeno.* <https://www.franzjimeno.es/index.php/8-blog/programacion/r/12-mapas-con-r>

[14] *Mapas y datos GADM. Mapas y datos espaciales para todos los países y sus subdivisiones administrativas.* <https://gadm.org/>

[15] *Scikit-learn. Biblioteca para aprendizaje automático en python.* <https://scikit-learn.org>

[16] *Calendario laboral Cataluña.* <https://www.elperiodico.com/es/economia>

9 ANEXO I. Descripción de los contaminantes

A continuación, a modo de información complementaria que puede resultar de interés se muestra algo más de detalle de las sustancias analizadas:

- Descripción
- Fuentes de emisión
- Aplicaciones en la actividad humana (en aquellos casos que así sea)
- Efectos para la salud

Dióxido de Azufre (SO_2)

Se trata de un gas tóxico liberado principalmente en la combustión de productos petrolíferos, principalmente el diésel, y el procesamiento de minerales tales como el carbón, el gas natural, etc, así como en procesos metalúrgicos, centrales eléctricas y calefacciones centrales.

En la naturaleza se encuentra en las proximidades de zonas volcánicas, siendo el principal causante de la lluvia ácida.

Desde el punto de vista de la salud presenta efectos tales como dificultad para respirar e inflamación de las vías respiratorias, irritación ocular, alteraciones psíquicas, edema pulmonar, paro cardíaco, colapso circulatorio, queratitis y problemas asociados a procesos de bronquitis y asma.

Además de ser altamente nocivo para la salud de las personas, lo es todavía más para las plantas, produciendo necrosis en los árboles, vegetación, deteriorando los suelos, materiales de construcción y monumentos históricos de piedra.



Monóxido de Carbono (CO)

El *Monóxido de Carbono* es un gas sin olor ni color, pero altamente tóxico. Puede causar súbitamente una enfermedad y la muerte.

Se encuentra en el humo de la combustión, siendo expulsado por vehículos, estufas, fogones de gas y sistemas de calefacción. El CO proveniente de estos humos puede acumularse en lugares que no tienen una buena ventilación pudiendo cualquier ser humano envenenarse al respirarlos.

Desde el punto de vista de la salud presenta efectos tales como dolor de cabeza, mareos, debilidad, náusea, vómitos, dolor en el pecho y confusión.

Óxidos de Nitrógeno (NO_x)

Se trata de un grupo de compuestos químicos gaseosos formados por la combinación de *Oxígeno* y *Nitrógeno*, siendo los más importantes el *Óxido Nítrico* o *Monóxido de Nitrógeno* (NO) y el *Dióxido de Nitrógeno* (NO₂).

- Monóxido de Nitrógeno (NO)

Es un gas que constituye uno de los contaminantes de la atmósfera, siendo uno de los agentes responsables de la lluvia ácida. Es altamente inestable en el aire ya que se oxida rápidamente en presencia de *Oxígeno* convirtiéndose en *Dióxido de Nitrógeno*. Por esta razón es considerado como un radical libre.

El *Monóxido de Nitrógeno* se genera en la atmósfera mediante reacciones químicas, a través de tormentas eléctricas y también debido a la actividad humana, por ejemplo, en las centrales térmicas y en las cámaras de combustión de los motores de explosión de los coches. Su efecto para con la radiación solar es doble. Mientras en la baja atmósfera contribuyen al calentamiento global, en el alta, lo hacen al oscurecimiento global.

Desde el punto de vista de la salud presenta efectos tales como disnea, broncoespasmo, dolor torácico, taquicardia, pueden existir leucocitosis y fiebre.

- Dióxido de Nitrógeno (NO₂)

Es un compuesto químico gaseoso de color marrón amarillento, tóxico e irritante, siendo uno de los principales contaminantes en las ciudades.

En la naturaleza se produce por los incendios forestales, las erupciones volcánicas y la descomposición de nitratos orgánicos, siendo el volumen total que se produce de forma natural infinitamente menor al que se produce por la actividad humana.

La mayor parte tiene su origen en la oxidación del monóxido de nitrógeno que se produce en la combustión de los motores de los vehículos. Es también un potenciador del material particulado, sobre todo de partículas finas MP_{2.5} que son las más perjudiciales. En su reacción con la luz ultravioleta del sol es un precursor de ozono troposférico (O³).



Desde el punto de vista de la salud presenta efectos relacionados con enfermedades de las vías respiratorias, tales como disminución de la capacidad pulmonar, bronquitis agudas, asma y se considera el culpable de los procesos alérgicos, sobre todo en niños. Se ha relacionado las exposiciones crónicas a bajo nivel con el enfisema pulmonar. Otros efectos menores son la irritación ocular y de las mucosas.

Partículas en Suspensión

Se trata de una sustancia o material particulado presente en la atmósfera de nuestras ciudades en forma sólida o líquida, tales como polvo, cenizas, hollín, partículas metálicas, cemento, polen, etc.

Se clasifican en dos grupos:

- Material particulado de diámetro igual o inferior a 2.5 micrómetros o PM_{2.5}
- Material particulado de diámetro igual o inferior a 10 micrómetros o PM₁₀

Cada tipo de partículas está compuesto de diferente material y puede provenir de diferentes fuentes. En el caso de las PM_{2.5}, su origen está principalmente en las emisiones de los vehículos diésel, mientras que las partículas PM₁₀ pueden tener en su composición un importante componente de tipo natural, como partículas de polvo procedente de las intrusiones de viento del norte de África (polvo sahariano).

Desde el punto de vista de la salud presenta efectos tales como enfermedades de tipo respiratorio, bronquitis y dolencias de tipo cardiovascular.

Ozono (O₃)

El *Ozono* es una molécula compuesta por tres átomos de oxígeno. Es altamente oxidante, inestable y se descompone rápidamente en oxígeno por efecto de la luz, calor y choques electrostáticos. Existen dos tipos de *Ozono*:

- *Ozono Estratosférico*. Se concentra en la estratosfera, lo que se denomina la capa de ozono. Tiene la capacidad de absorber muy eficazmente la radiación ultravioleta procedente del sol.
- *Ozono Troposférico*. Se genera mediante reacciones químicas a partir de los óxidos de nitrógeno y materia orgánica volátil resultante de la quema de combustibles fósiles en ciudades y zonas industriales. El *Ozono* resultante es un poderoso oxidante, que aparte de actuar como gas de efecto invernadero, en elevadas concentraciones, afecta muy negativamente en la salud, presentando efectos tales como irritación de los ojos y las vías respiratorias, alergias, dolores de cabeza y daños orgánicos más graves, además del impacto que tiene en las plantas, alterando su actividad fotosintética.



Arsénico (As)

El *Arsénico* (As) es un mineral altamente tóxico que se encuentra de manera natural en la corteza terrestre y su dispersión en el ambiente se puede hacer a través de aguas subterráneas, minerales y procesos geotérmicos. También es liberado a través de volcanes, por erosión de depósitos minerales y por procesos comerciales e industriales, por su uso en herbicidas, medicamentos, etc.

Desde el punto de vista de la salud afecta prácticamente a todos los aparatos y sistemas del cuerpo, puesto que interfiere negativamente tanto en reacciones enzimáticas como en la respiración celular.

Plomo (Pb)

El *Plomo* es un metal pesado que existe de forma natural en la corteza terrestre. Presenta las mayores concentraciones como consecuencia de la actividad humana, siendo la combustión del petróleo, los residuos sólidos y los procesos industriales algunas de ellas.

Desde el punto de vista de la salud presenta efectos tales como perturbación de la biosíntesis de hemoglobina y anemia, incremento de la presión sanguínea, daño a los riñones, abortos, perturbación del sistema nervioso, daño al cerebro, disminución de la fertilidad del hombre a través del daño en el esperma, disminución de las habilidades de aprendizaje de los niños y perturbación en el comportamiento de los niños, como es agresión, comportamiento impulsivo e hipersensibilidad.

Benzoapireno (BAP)

Los *Benzopirenos* son un grupo de compuestos químicos que se forman por la combustión incompleta de materia orgánica. En países desarrollados, debido al aumento de la actividad industrial, se liberan al medio ambiente durante la combustión de madera, petróleo, aceites, carbón, basuras, tabaco y alimentos como carne y pescado.

Desde el punto de vista de la salud el *Benzopireno* puede causar erupciones en la piel, sensación de quemazón, cambios en el color de la piel, verrugas y bronquitis, así como procesos cancerígenos.

Cadmio (Cd)

El *Cadmio* tiene relación estrecha con el *Zinc*, con el que se encuentra asociado en la naturaleza. Se encuentra mayoritariamente en la corteza terrestre siendo liberado en el ambiente de forma natural a través de la descomposición de rocas y a través de fuegos forestales y volcanes. El resto es liberado por la actividad humana, en la producción de fertilizantes, manufacturación, la quema de residuos urbanos y de combustibles fósiles entre otros.



Desde el punto de vista de la salud puede causar diarreas, dolor de estómago y vómitos severos, fractura de huesos, fallos en la reproducción y posibilidad incluso de infertilidad, daño al sistema nervioso central, daño al sistema inmune, desórdenes psicológicos y posible daño en el ADN o desarrollo de cáncer.

Benceno (C₆H₆)

El *Benceno* es un líquido muy tóxico, incoloro e inflamable que se evapora rápidamente cuando se expone al aire. Se forma a partir de procesos naturales, como los volcanes y los incendios forestales, pero la mayor exposición al benceno es el resultado de las actividades humanas. Es parte natural del petróleo crudo y la gasolina, siendo utilizado en procesos industriales, para fabricar productos químicos y plásticos, lubricantes, cauchos, fibras sintéticas, colorantes, asfalto, en equipos de impresión y pintura.

Desde el punto de vista de la salud, el *Benceno* presenta efectos tales como dolores de cabeza, mareos, confusión, somnolencia, temblores, inconsciencia y desarrollo de cáncer, pudiendo incluso provocar la muerte a niveles altos.

Níquel (Ni)

El *Níquel* aparece de forma natural en los meteoritos y se encuentra en el núcleo de la Tierra. Está presente en el ambiente a muy pequeños niveles y su uso principal es como ingrediente del acero y otros productos metálicos. Es liberado al aire por plantas de energía e incineradoras de basura y, principalmente, también a través del tabaco.

Desde el punto de vista de la salud el *Níquel* presenta efectos tales como elevadas probabilidades de desarrollar cáncer de pulmón, nariz, laringe y próstata, enfermedades y mareos, embolia de pulmón, fallos respiratorios, defectos de nacimiento, asma, bronquitis crónica, reacciones alérgicas y desórdenes del corazón.

10 ANEXO II. Evaluación de la Calidad el Aire

Para evaluar la calidad del aire, la legislación de la Comisión Europea considera diferentes medidas que ayudan a establecer unos umbrales en relación a los valores permitidos de contaminación de los diferentes gases/metales de efecto invernadero, considerando diversos objetivos de calidad:

- *Valores límite* para la protección de la salud
- *Valores objetivo* y objetivo a largo para la protección de la salud.
- *Niveles críticos* para la protección de la vegetación.

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



El *valor objetivo* es el nivel de un contaminante que deberá alcanzarse, en la medida de lo posible, en un momento determinado para evitar, prevenir o reducir los efectos nocivos sobre la salud humana, el medio ambiente en su conjunto y demás bienes de cualquier naturaleza.

A su vez, el *objetivo a largo plazo* es el nivel de un contaminante que debe alcanzarse a largo plazo, salvo cuando ello no sea posible con el uso de medidas proporcionadas, con el objetivo de proteger eficazmente la salud humana, el medio ambiente en su conjunto y demás bienes de cualquier naturaleza.

Finalmente, el *nivel crítico* es aquel fijado con arreglo a conocimientos científicos por encima del cual pueden producirse efectos nocivos para algunos receptores como las plantas, árboles o ecosistemas naturales, pero no para el ser humano.

A continuación, se indican los valores, por sustancia, que fija la legislación para dichos niveles.

Valores límite para la protección de la salud

Contaminante	Período de promedio	Valor límite	Fecha de cumplimiento	Umbral de alerta
SO ₂	Horario	350 µg/m ³ (24 superaciones como máximo al año)	01/01/2005	500 µg/m ³ (en 3 horas)
	Diario	125 µg/m ³ (3 superaciones como máximo al año)	01/01/2005	--
NO ₂	Horario	200 µg/m ³ (18 superaciones como máximo al año)	01/01/2010	400 µg/m ³ (en 3 horas)
	Anual	40 µg/m ³	01/01/2010	--
PM ₁₀	Diario	50 µg/m ³ (35 superaciones como máximo al año)	01/01/2005	--
	Anual	40 µg/m ³	01/01/2005	--
Pb	Anual	0,5 µg/m ³	01/01/2005	--
C ₆ H ₆	Anual	5 µg/m ³	01/01/2010	--
CO	Máximo diario de las medias	10 mg/m ³	01/01/2005	--



	móviles octohorarias			
PM _{2.5}	Anual	25 µg/m ³	01/01/2015	—

Valores objetivo para la protección de la salud

Contaminante	Período de promedio	Valor objetivo	Objetivo a largo plazo	Fecha de cumplimiento	Umbral de información	Umbral de alerta
PM _{2.5}	Anual	25 µg/m ³	—	01/01/2010	—	—
As	Anual	6 ng/m ³	—	01/01/2013	—	--
Cd	Anual	5 ng/m ³	—	01/01/2013	—	—
Ni	Anual	20 ng/m ³	—	01/01/2013	—	—
B(a)P	Anual	1 ng/m ³	—	01/01/2013	—	—
O ₃	Horario	—	—	01/01/2004	180 µg/m ³	240 µg/m ³ (en 3h)
	Máximo diario de las medias móviles octohorarias	120 µg/m ³ (25 superaciones como máximo, en un promedio de 3 años)	--	01/01/2010 (periodo trianual 2010-2012)	—	—
		—	120 µg/m ³	No definida	—	--

Niveles críticos para la protección de la vegetación

Contaminante	Período de promedio	Nivel crítico	Valor objetivo	Objetivo a largo plazo	Fecha de cumplimiento
SO ₂	Anual e invierno (1-octubre a 31-marzo)	20 µg/m ³	—	—	11/06/2008



No _x	Anual 30 µg/m ³ (expresado como NO ₂)	–	–	–	11/06/2008
O ₃	AOT40 ⁶ a partir de valores horarios, de mayo a julio	–	18000 µg/m ³ (promedio en un periodo de 5 años)	–	01/01/2010 (periodo quinquenal 2010-2014))
		–	–	6000 µg/m ³ h	No definida

La metodología de evaluación de los contaminantes indicados establece unos límites o *umbrales superior e inferior* de evaluación contemplados en la legislación con los siguientes valores:

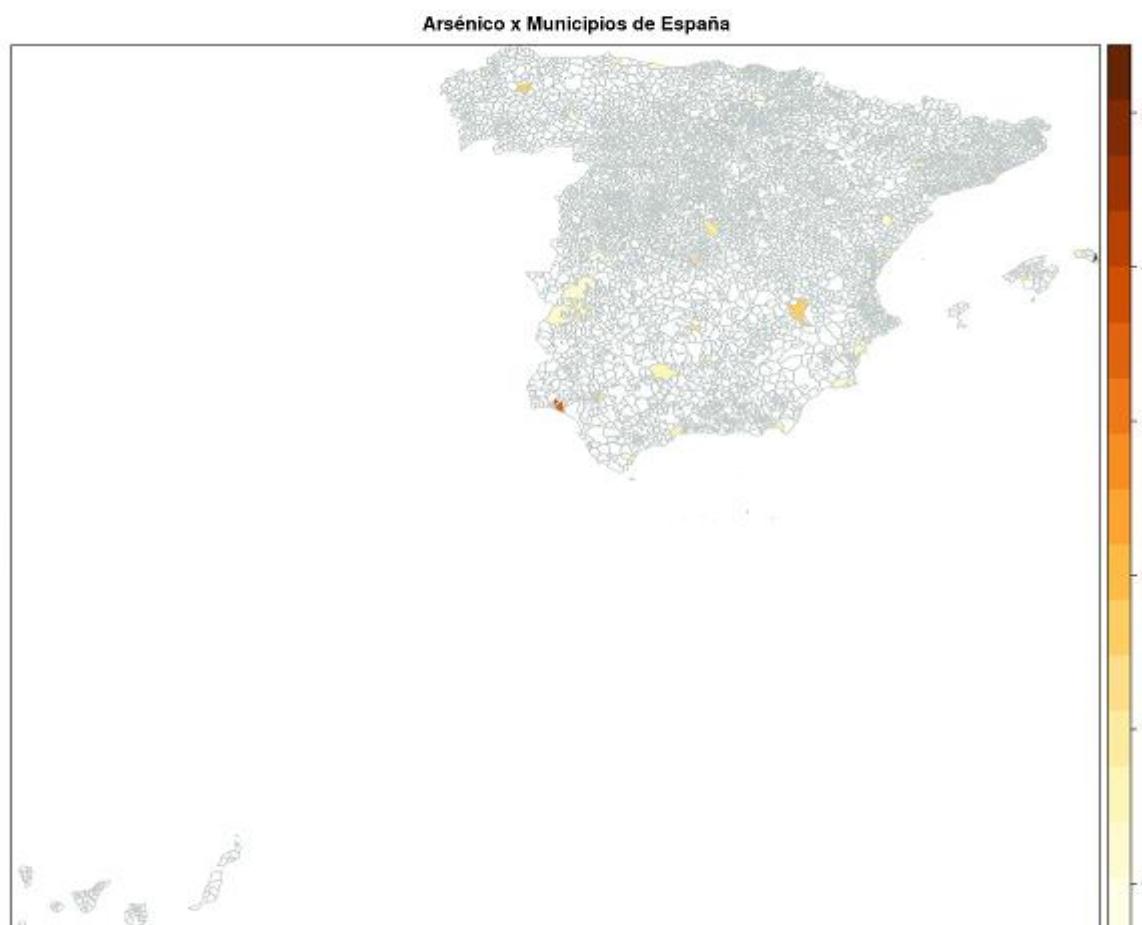
<u>Umbrales superiores e inferiores de evaluación</u>			
	Contaminante	Umbral superior de evaluación	Umbral inferior de evaluación
SO ₂	Protección de la salud	60% de VLD (75 µg/m ³ no más de 3 ocasiones/año)	40% de VLD (50 µg/m ³ no más de 3 ocasiones/año)
	Protección de la vegetación	60% del nivel crítico de invierno (12 µg/m ³ , del nivel crítico de invierno)	40% del nivel crítico de invierno (8 µg/m ³ del nivel crítico de invierno)
NO ₂	Valor límite horario para la protección de la salud humana	70% del VLH (140 µg/m ³ no más de 18 ocasiones/año)	50% del VLH (100 µg/m ³ no más de 18 ocasiones/año)
	Valor límite anual para la protección de la salud humana	80% del VLA (32 µg/m ³)	65% del VLA (26 µg/m ³)
No _x	Nivel crítico anual para la protección de la vegetación y los ecosistemas	80% del nivel crítico (24 µg/m ³ como NO ₂)	65% del nivel crítico (19,5 µg/m ³ como NO ₂)
PM ₁₀	Media diaria	70% del VLD (35 µg/m ³ no más de 35 ocasiones/año)	50% del VLD (25 µg/m ³ no más de 35 ocasiones/año)
	Media anual	70% del VLA (28 µg/m ³)	50% del VLA (20 µg/m ³)
PM _{2,5}	Media anual	70% del VLA (17 µg/m ³)	50% del VLA (12 µg/m ³)
Pb	Media anual	70% del VLA (0,35 µg/m ³)	50% del VLA (0,25 µg/m ³)



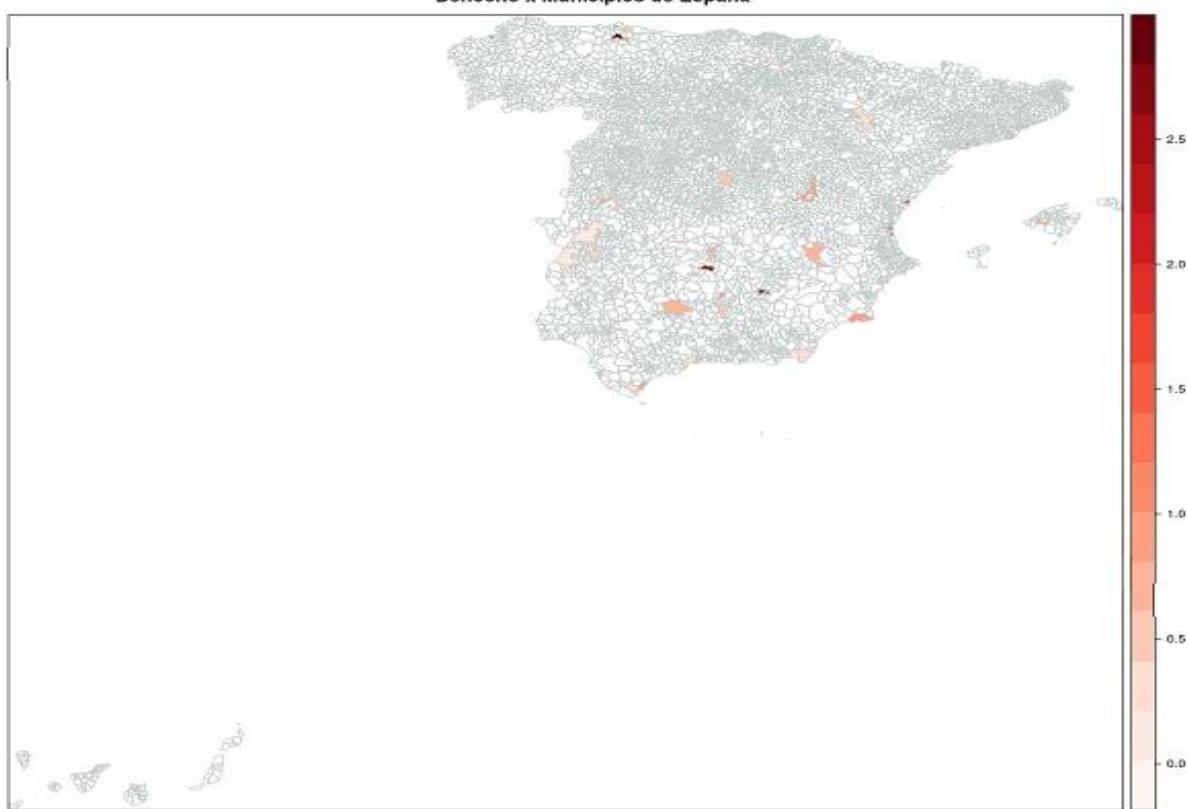
C ₆ H ₆	Media anual	70% del VLA (3,5 µg/m ³)	40% del VLA (2 µg/m ³)
CO	Promedio de períodos de 8 horas	70% del VL (7 µg/m ³)	50% del VL (5 mg/m ³)
As	Media anual	60% del VO (3,6 ng/m ³)	40% del VO (2,4 ng/m ³)
Cd	Media anual	60% del VO (3 ng/m ³)	40% del VO (2 ng/m ³)
Ni	Media anual	70% del VO (14 ng/m ³)	50% del VO (10 ng/m ³)
B(a)P	Media anual	60% del VO (0,6 ng/m ³)	40% del VO (0,4 ng/m ³)

11 ANEXO III. Gráficos de contaminantes por ubicación geográfica

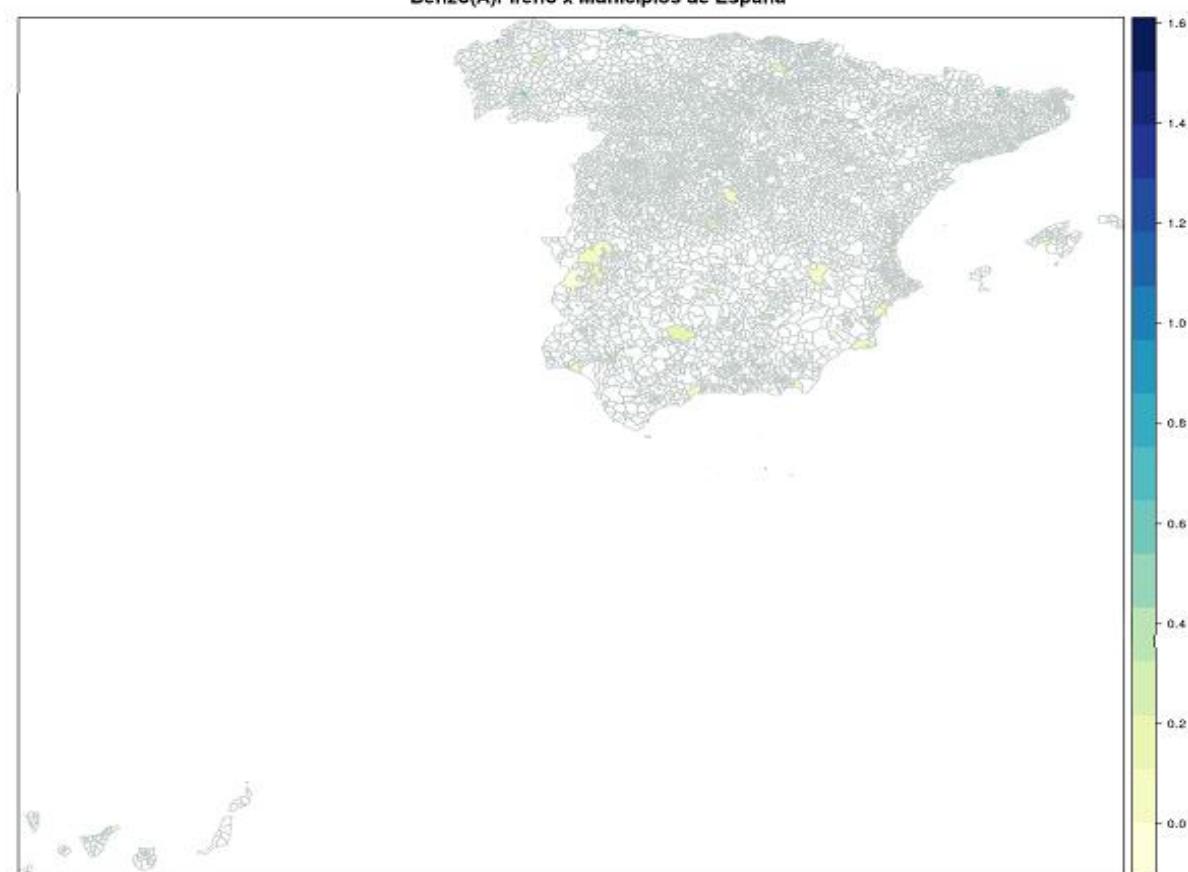
En este apartado se adjuntan gráficos de otros contaminantes por ubicación geográfica en España y las islas no incluidos en el apartado de análisis descriptivo.



Benceno x Municipios de España

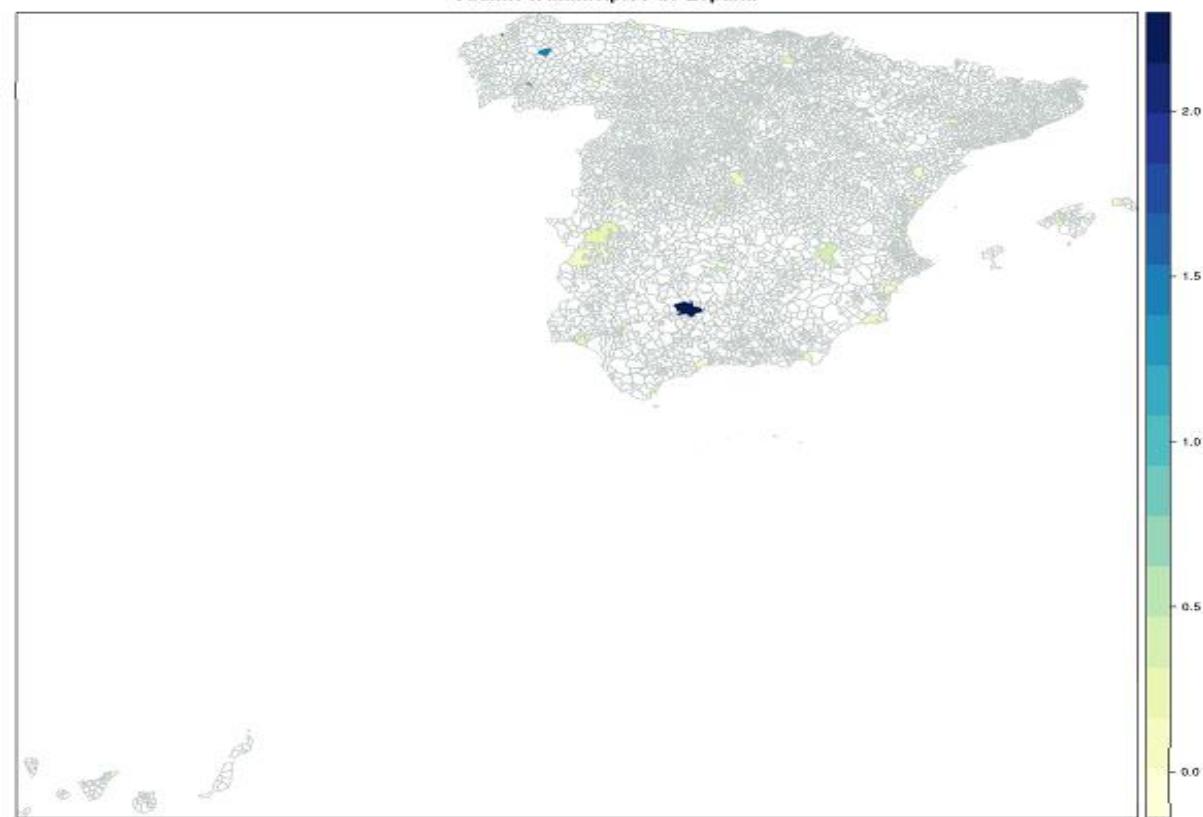


Benzo(a)Pireno x Municipios de España

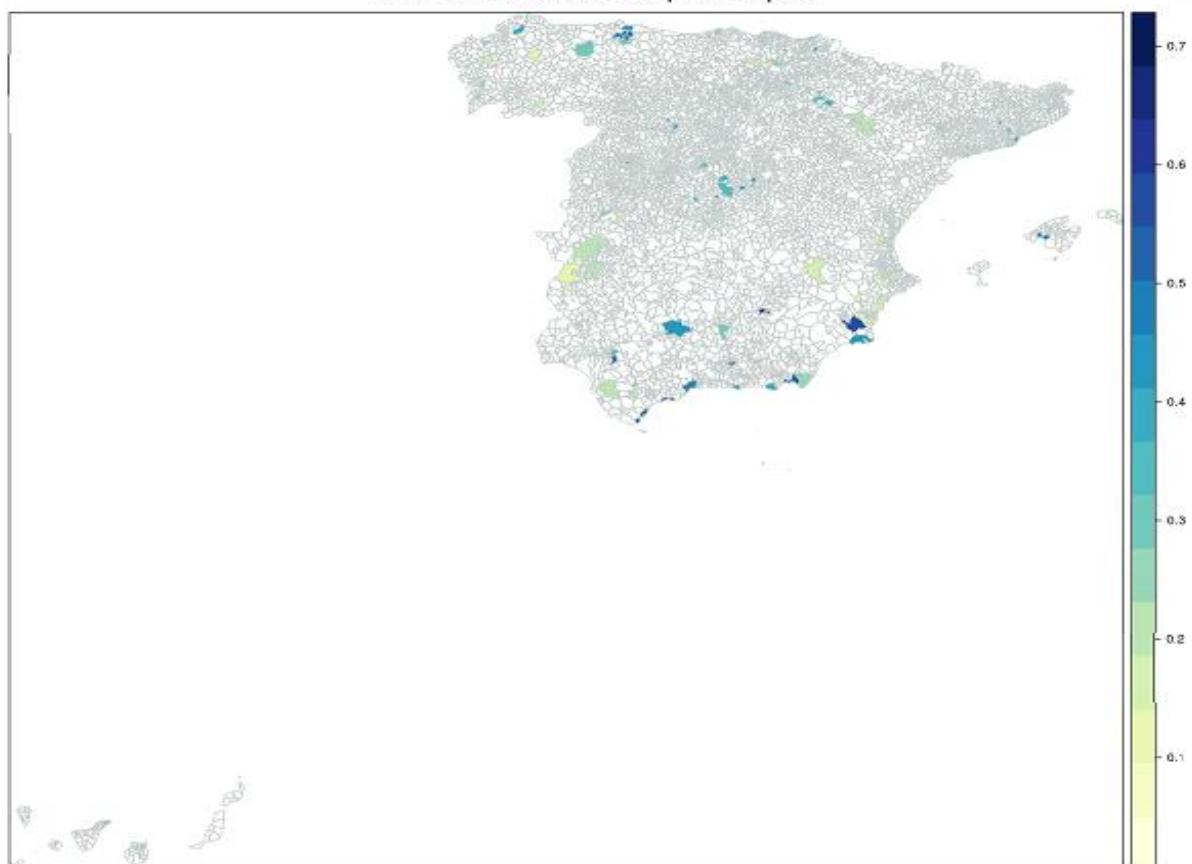


Evaluación sobre la calidad del aire en España

Cadmio x Municipios de España

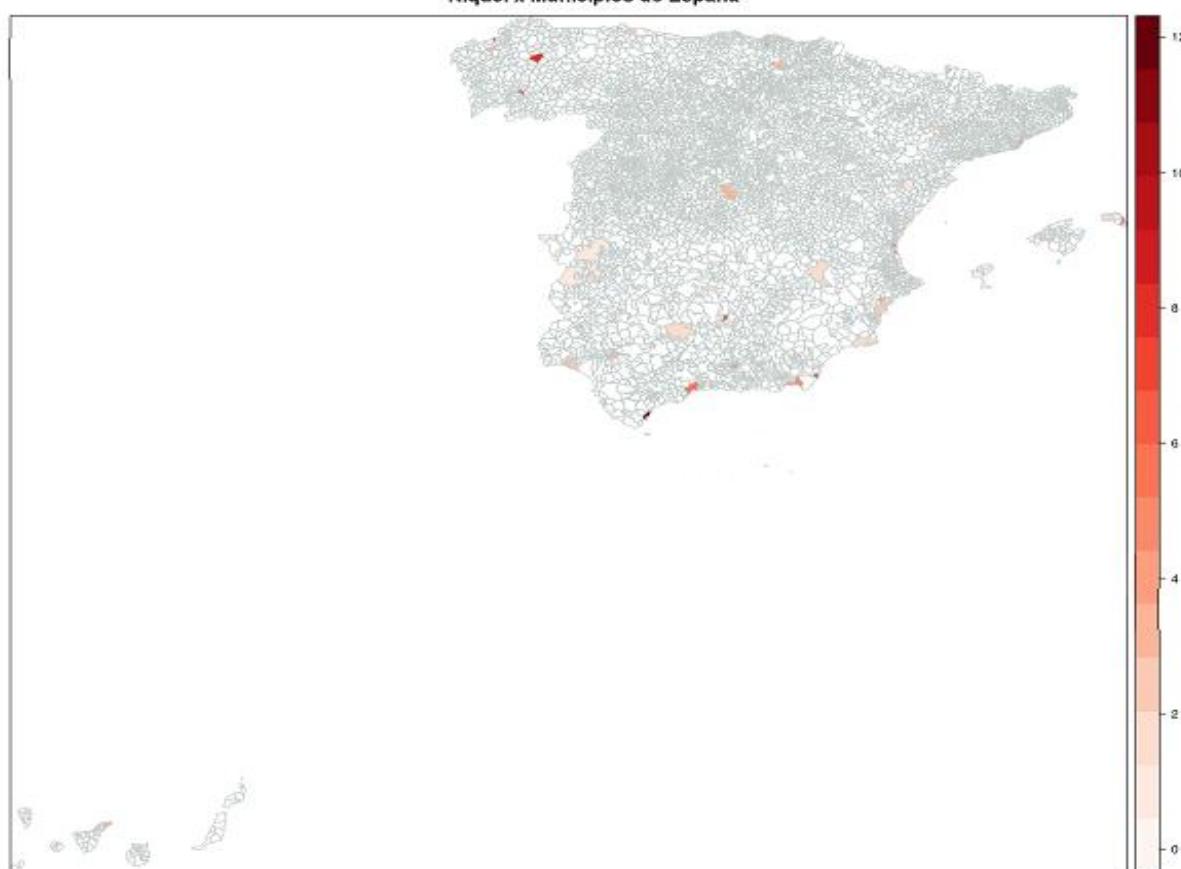


Monóxido de Carbono x Municipios de España

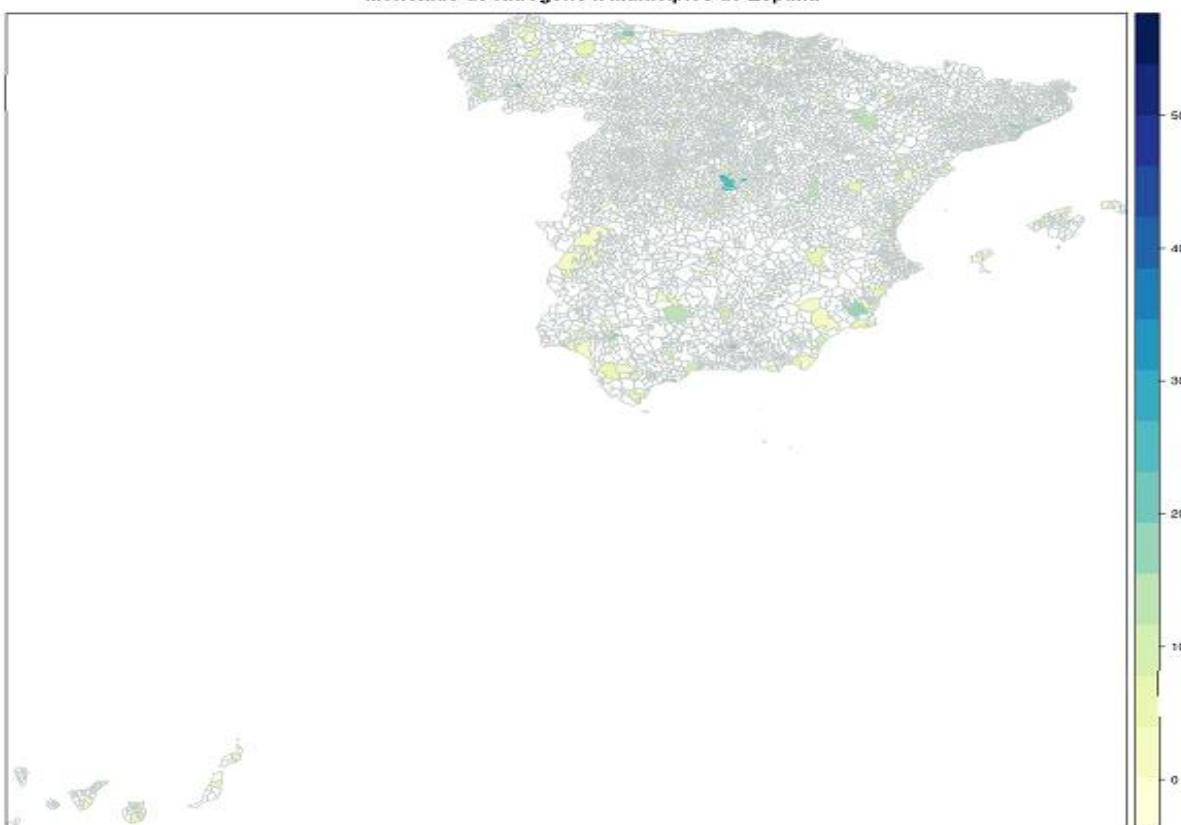


Evaluación sobre la calidad del aire en España

Níquel x Municipios de España



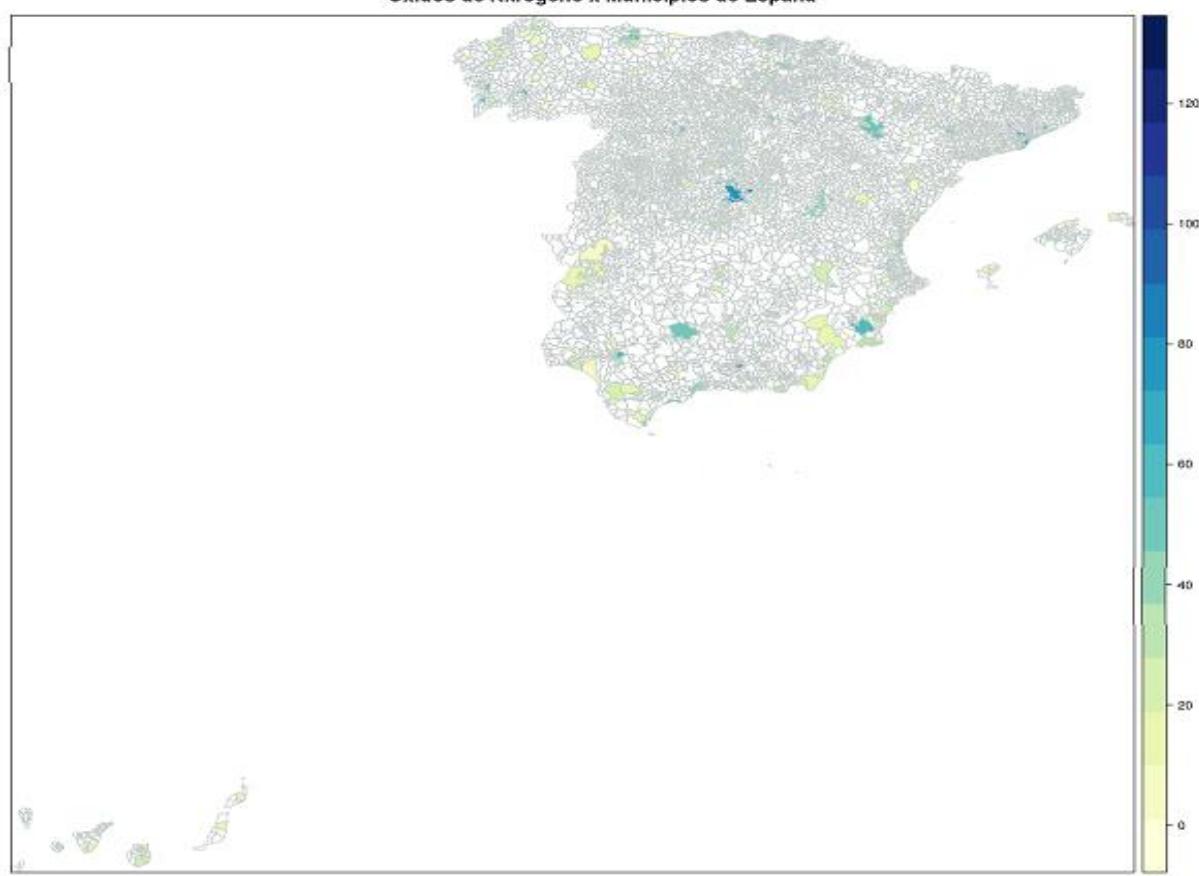
Monóxido de Nitrógeno x Municipios de España



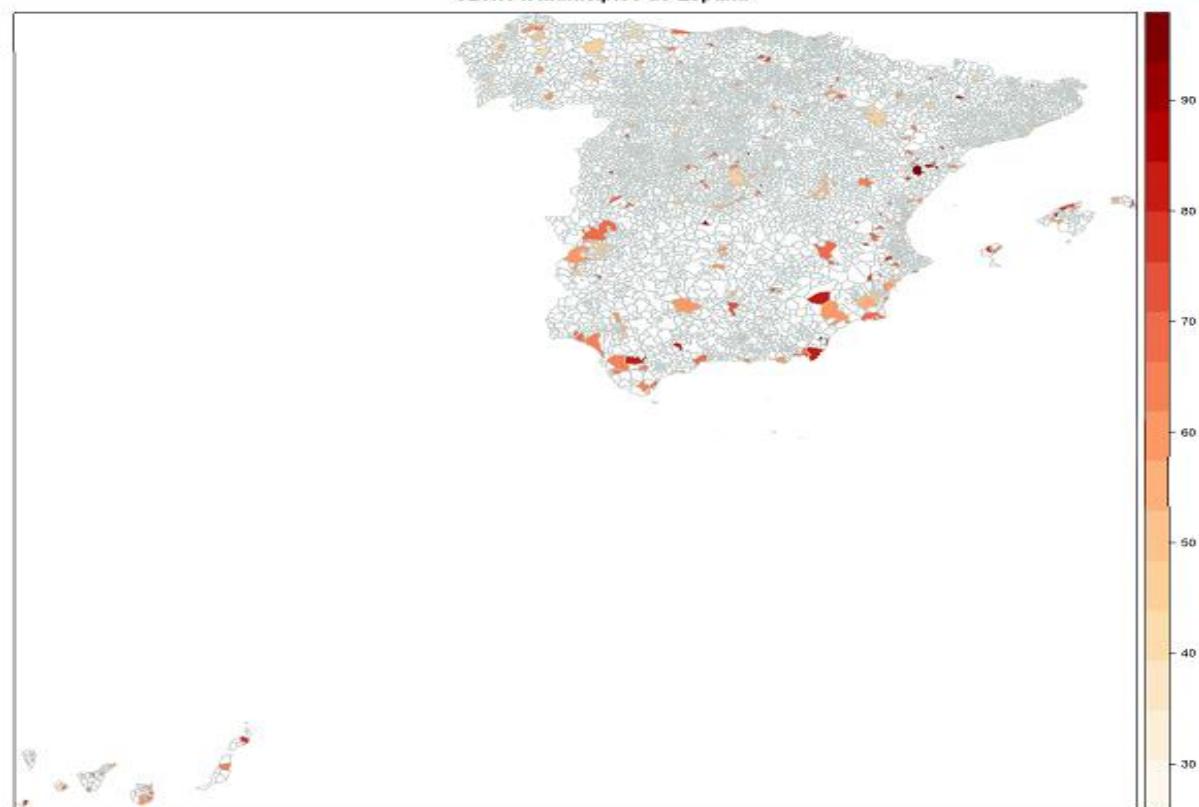
100



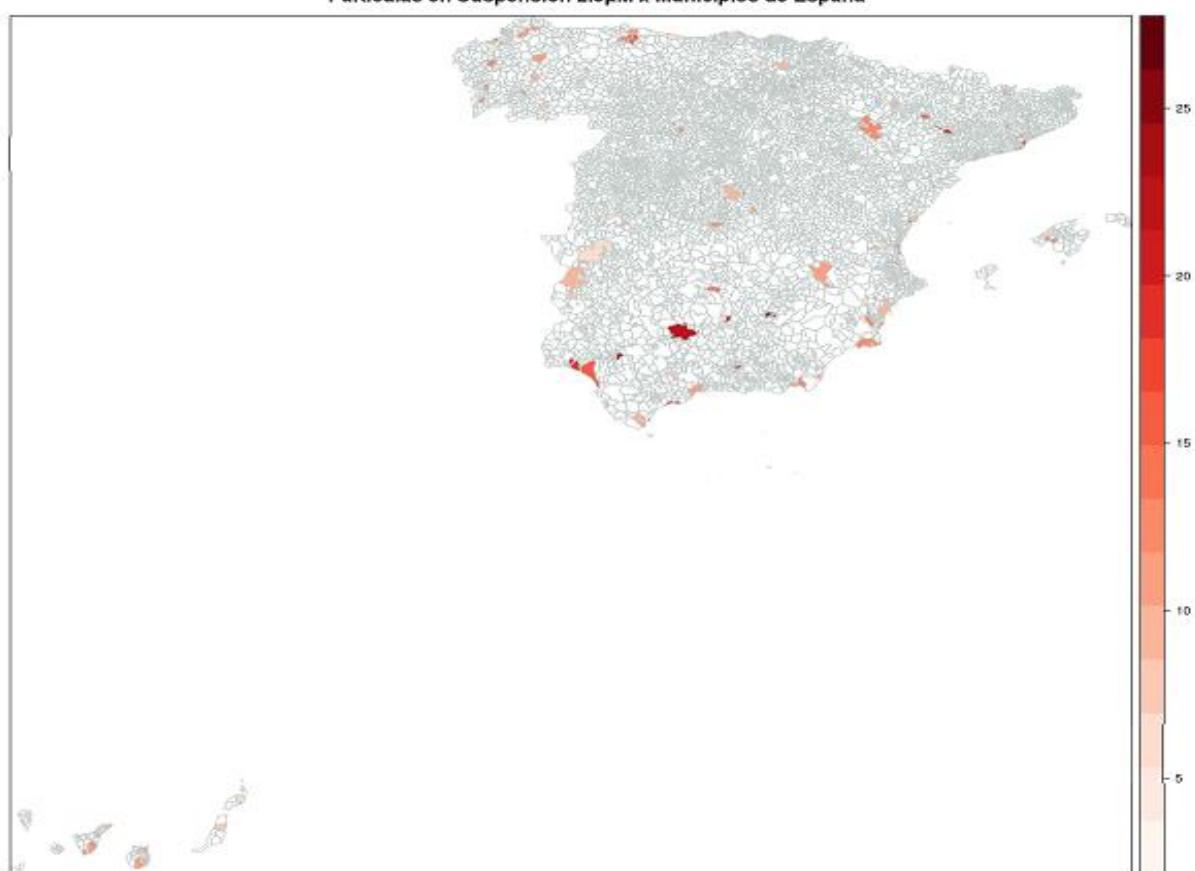
Óxidos de Nitrógeno x Municipios de España



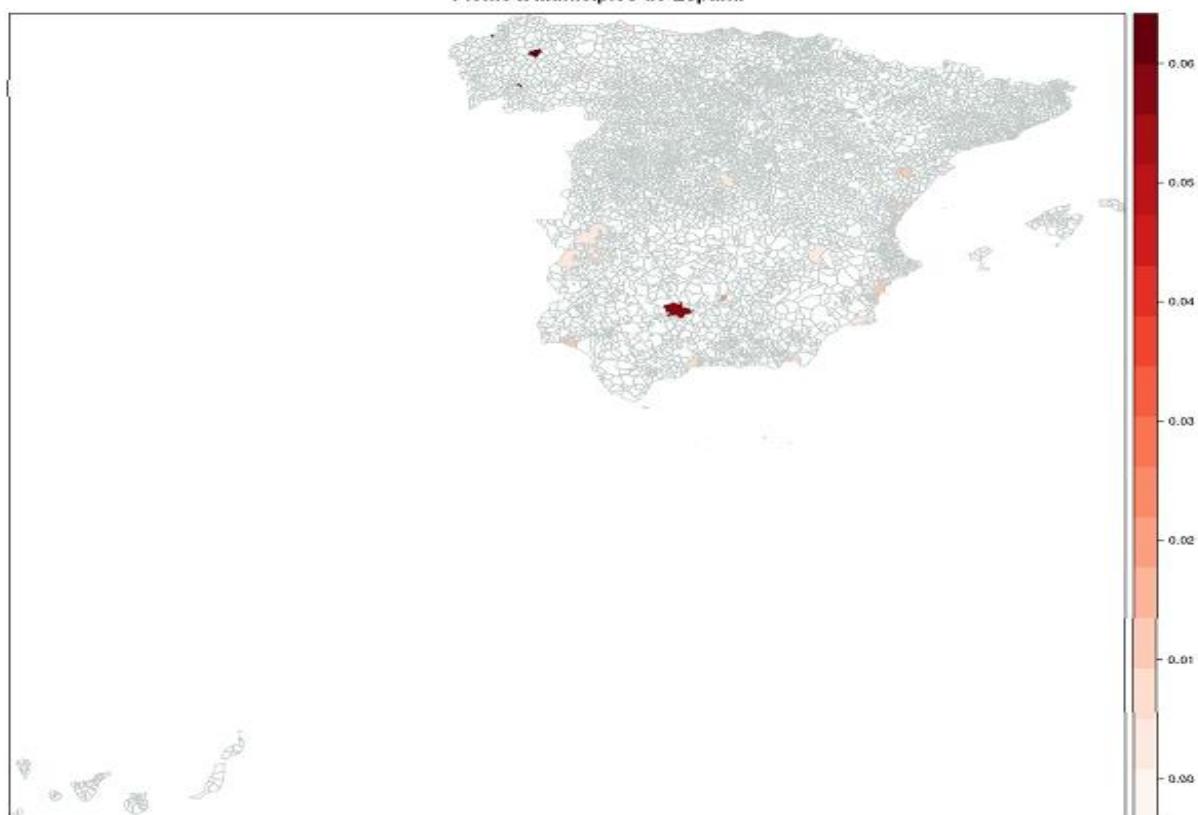
Ozono x Municipios de España



Partículas en Suspensión 2.5 μ M x Municipios de España



Pbomo x Municipios de España



12 ANEXO IV. Trabajos relacionados

Existen multitud de trabajos relacionados y referencias de este tipo de estudios. A continuación, indico un breve resumen de los que parecen interesantes.

- *Análisis de la Calidad del Aire en España. Evolución 2001 – 2012, elaborado por el Ministerio de Agricultura, Alimentación y Medio Ambiente*

<https://es.scribd.com/document/375629687/Analisis-Calidad-Aire-Espana-2001-2012-Tcm7-311112>

- *Evaluación de la Calidad del Aire en España. Año 2017, elaborado por el Ministerio para la Transición Ecológica*

<https://www.miteco.gob.es/es/prensa/ultimas-noticias/La-calidad-del-aire-en-Espa%C3%B1a-en-2017-baja-levemente-con-respecto-al-a%C3%B1o-anterior/tcm:30-481677>

- *Plan Nacional de Calidad del Aire. 2017 – 2019 (Plan Aire II), elaborado por el Ministerio de Agricultura, Alimentación y Medio Ambiente*

https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/planaire2017-2019_tcm30-436347.pdf

Dentro de estos documentos existen numerosas referencias a:

- *Bases científico-técnicas para un Plan Nacional de Mejora de la Calidad del Aire*
- *Estudio y evaluación de la contaminación atmosférica por sustancia contaminante específica*
- *Anuarios Estadísticos del Ministerio de Agricultura y Medio Ambiente*
- *Histórico de la calidad del aire en España*
- *Informes anuales de evaluación de la calidad del aire para la Comisión Europea*
- *Legislación de calidad del aire de la Comisión Europea*
- *Políticas medioambientales de UNECE (United Nations Economic Commission for Europe)*
- *Registro Estatal de Emisiones y Fuentes Contaminantes*



- *Tratados sobre contaminación atmosférica UNECE (United Nations Economic Commission for Europe)*

También existe información recopilada en bases de datos, foros, etc:

- *Base de datos europea de calidad del aire: AIRBASE*
- *Foro para la modelización de la calidad del aire: FAIRMODE*
- *Red Temática de Modelización de la Contaminación Atmosférica: RETEMCA*

así como herramientas de predicción:

- *Sistema de pronóstico de la calidad del aire operativo para España: CALIOPE*

