PROYECTO TFM

"Análisis de dependencia de la contaminación atmosférica con la tasa de desempleo en España. Visualización con series temporales"

Gonzalo Gómez Rivas

Junio - 2018

ÍNDICE

1.- Introducción

2.- Obtención de datos para el proyecto

3.- Tipología de los datos a tratar

4.- Metodología y manual de uso

5.- Principales resultados

6.- Conclusiones

Anexo

1.- Introducción

El presente trabajo se planteó con varias motivaciones, la primera de ellas era intentar probar mediante análisis de datos obtenidos si las mediciones de contaminación atmosférica en España se ven influenciadas por la tasa de empleo, algo que a priori parece evidente, pero que por la información publicada hasta el momento no es fácil de comprobar a nivel de país (sí que existe más información de estudios en grandes ciudades)

Se planteó también este trabajo como un reto por conseguir un proyecto completo utilizando datos abiertos de diferentes organismos nacionales. Por experiencias previas ya se sabía que era bastante complejo por la escasa implicación de las diversas administraciones públicas en liberar datos de manera sencilla y transparente.

Para ejecutar el proyecto se realizó una búsqueda bastante intensa del estado del arte, no encontrando investigaciones completas acerca de la concentración de diferentes sustancias químicas contaminantes, su relación con la tasa de desempleo y la variabilidad temporal.
Es cierto que existen muchas publicaciones de las mediciones de calidad del aire y contaminantes en diferentes años, pero sin centrarse en la relación arriba mencionada (en el anexo final se citan algunos de estos artículos).

El estudio se lleva a cabo con el manejo de series temporales, con lo que puede realizarse un análisis de visual y por métodos de regresión lineal. Esto permite de manera sencilla comprobar hasta qué punto es o no correcto el planteamiento de este trabajo.

2.- Obtención de datos para el proyecto

Los datos para este trabajo han sido obtenidos principalmente de dos fuentes:

- Instituto Nacional de Estadística (INE)

- Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente (MAPAMA)

Igualmente se han accedido a otras fuentes de información para la interpretación de algunos datos, especialmente a las mediciones de contaminantes. Estas fuentes han sido:

- Instituto de Salud Carlos III

- Organización Mundial de la Salud (OMS)

- Agencia Europea de Medio Ambiente (EEA)

- División de Estadísticas de Naciones Unidas (UNSTATS)

(En el anexo se citan otros enlaces consultados)

3.- Tipología de los datos a tratar

1- Datos tasa de desempleo en España

Se realizó multitud de búsquedas en la base de datos de INEbase, y estudio de la Encuesta de Población Activa (EPA) que el INE publica trimestralmente.

De estos datos los que nos interesaban para la investigación eran:

- tasa de desempleo nacional desde el año 2002 (el primer año que se podía consultar) hasta la actualidad

- tasa de desempleo nacional por sector de actividad. Existen 21 ramas de actividades económicas. Se evaluaron aquellas que tanto por el porcentaje de paro en estos años, así como por la capacidad, a priori, de emitir sustancias contaminantes al aire pudieran ser de interés para el posterior análisis exploratorio de datos y su uso con series temporales.

El INE proporciona un sistema completo para adquisición de datos en formato .csv o Excel aunque las tablas de datos en crudo que se descargan necesitan de un tratamiento previo al análisis de datos.

2- ***Datos de contaminación atmosférica***

Los datos de contaminación atmosférica proceden de fuentes oficiales del MAPAMA específicamente de los publicados en:

datos de calidad del aire 2001-2016

Estos datos proceden de 1198 estaciones medidoras de la contaminación repartidas por todo el país.

La estructura de datos es la siguiente:

Fichero .csv por contaminante y por año. Cada fichero recoge las mediciones diarias durante todo el año en las 1198 estaciones de un contaminante específico.

El delimitador de dato es punto y coma.
Separador de decimales: punto
Sólo se dan los datos válidos
El valor puede estar vacío por dos razones:

- Dato no válido
- No se ha medido en ese periodo

Por desgracia, no fue hasta su descarga, descompresión de carpetas y estudio de las mismas cuando se evaluó la calidad del material. Nos encontramos con varios problemas:

- Pésimos datos por la cantidad de mediciones nulas diarias de cada contaminante y estación
- Algunos datos interpretados como numéricos y otros como textos.
- Carencia de multitud de contaminantes químicos que se utilizan para la mediciones de la contaminación atmosférica. Estos contaminantes están así reconocidos en el Real Decreto 102/2011, de 28 de enero (que establece la evaluación de la calidad del aire ambiente), como son dióxido de azufre (**SO_2**), dióxido de nitrógeno y óxidos de nitrógeno (**NO_2, NO_X**) , benceno, monóxido de carbono (**CO**) y ozono (**O_3**) simplemente no aparecen en los ficheros públicos de contaminantes del Ministerio.
- Hasta el año 2009 el Ministerio sólo recoge (al menos públicamente) dos sustancias: Plomo (Pb) y partículas en suspensión de 10 micras (PM10) , no siendo hasta el año 2009 cuando recogen otros cinco contaminantes más: Niquel (Ni), Cadmio (Cd), Arsénico (As), Benzenopireno y partículas contaminantes de 2.5 micras (PM25)
- Falta completa de colaboración del Ministerio a las consultas realizadas. Ninguna de las tres consultas que se le hizo tuvieron respuesta.

A causa de todo ello, el proyecto arrastraba un problema claro de deficiencia de datos. Se optó entonces por centrarse en el análisis con las partículas en suspensión de 10 micras (PM10) , uno de lo componentes más importantes en casi todos los estudios que se suelen realizar de contaminación atmosférica.

De dichos datos se cuenta con una serie temporal diaria desde el año 2002 hasta la actualidad, por lo que se cruzará con la tasa de paro mensual de ese periodo.

Nota: En el anexo se encuentran los diferentes enlaces donde pueden consultarse los datos en “crudo” que se han utilizado para el estudio.

4.- Metodología

El proyecto ha sido realizado íntegramente con Python y el uso de las siguientes librerías:

Pandas, Matplotlib, Numpy, Seaborn , Scikit-learn, Plotly

Proceso:

1.- Se realiza una primera aproximación a los datos en crudo. En el siguiente notebook visualizamos algunos de estos datos para ver como tratarlos posteriormente:

Notebook: “Limpieza y AED PM_10.ipynb”

2.- En el siguiente notebook se realizan los cálculos y algoritmos necesarios para obtener agregados y organizados todos los contaminantes que el Ministerio de Medio Ambiente pone a disposición pública:

Notebook: “Algoritmo para todos los contaminantes.ipynb”

3.- Análisis exploratorio de datos. En este notebook se realiza un breve análisis exploratorio, en el que como explica más adelante esta memoria, apenas existe correlación entre los distintos contaminantes, y con la tasa de paro, a excepción de las partículas PM10 que son aquellas que por su importancia nos enfocaremos en su estudio.

Notebook: “Análisis exploratorio.ipynb”

4.- Método estadístico / regresión. Una vez preparados y analizados los datos, se procede a estudiar con un modelo de regresión lineal los datos obtenidos. Se realiza un pequeño modelo predictivo, y métricas de evaluación de regresión. Utilizamos para ello la librería Sklearn

Notebook: “Estudio de Regresión Lineal.ipynb”

5.- Pasamos a la visualización de datos a través de series temporales, donde de manera bastante intuitiva se aprecia la relación de la contaminación con la tasa de paro según un periodo temporal. Para ello hacemos uso intensivo de la librería Matplotlib

Notebook: “Visualización.ipynb”

6.- Por último todos los resultados obtenidos son agregados y visualizados en un dashboard interactivo creado con la librería Plotly.

Notebook: “Dashboard final.ipynb”

Nota: Todos los datos necesarios que los archivos ejecutan se encuentra disponible en la carpeta *“datos”*

5.- Principales resultados

Para poder analizar los resultados es necesario hacer un breve inciso sobre algunos de los contaminantes medidos:

- El níquel (Ni) es liberado al aire por las plantas de energía y las incineradoras de basuras. Este se depositará en el suelo o caerá después de reaccionar con las gotas de lluvia. Usualmente lleva un largo periodo de tiempo para que el níquel sea eliminado del aire.

- Cadmio (Cd). Casi todo el que se produce es obtenido como subproducto de la fundición y refinamiento de los minerales de zinc en la industria.

- Arsénico (As). Puede presentarse en forma de partículas en suspensión en el aire como producto de las emisiones de las industrias que utilizan procesos a alta temperatura, como la producción de energía quemando carbón o los hornos de fundición.

- las partículas PM10 se pueden definir como aquellas partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento ó polen, dispersas en la atmósfera, y cuyo diámetro varía entre 2,5 y 10 µm (1 micrómetro corresponde la milésima parte de 1 milímetro). Están formadas principalmente por compuestos inorgánicos como silicatos y aluminatos, metales pesados entre otros, y material orgánico asociado a partículas de carbono (hollín). Se caracterizan por poseer un pH básico debido a la combustión no controlada de materiales.
Un porcentaje importante de estas partículas procede del polvo resuspendido existente en la atmósfera. La industria, la construcción, el comercio y el transporte rodado con un representan otros focos de contaminación de especial relevancia.

Interpretación:

Se observó que de las mediciones de los contaminantes atmosféricos Ni, Cd, As, Benzenopireno, que proporcionaba el Ministerio se producían resultados poco esperados. Por ejemplo que apenas existía correlación con una mayor o menor tasa de desempleo en los sectores de la construcción, transporte y manufactura; lo que a priori es extraño pues algunos de estos contaminantes proceden de su liberación a la atmósfera en procesos industriales.
A la vista de estos resultados, investigamos las posibles causas, descubriendo que:

- La directiva europea 2004/107/CE establece una normativa relativa al arsénico, el cadmio, el mercurio, el níquel y los hidrocarburos aromáticos policíclicos en el aire ambiente. En dicha ley se establece que a partir del año 2012 los países restringirán la cantidad de estos contaminantes a unos límites bastante inferiores a los del año 2004.
Teniendo en cuenta que para casi todos estos contaminantes sólo disponíamos de datos a partir del año 2009, es claro que la industria había ido poco a poco reduciendo, y en algunos casos eliminando estos componentes de sus procesos industriales. Conociendo además que la industria incluye las cementeras y sectores afines muy relacionados así mismo con la construcción.

Así se aprecia en el análisis exploratorio que estos contaminantes no tienen relación directa con el problema a estudiar, por lo que se descartan en el estudio posterior.

Como se sospechaba, las partículas PM10 sí tenían una correlación inversa bastante fuerte con la tasa de desempleo. Esto es, se observa que cuanto mayor tasa de desempleo menor contaminación por dichas partículas.
Es muy interesante comprobar el periodo de la gran crisis económica de los años 2009-2013 en el que la tasa de desempleo crece abruptamente y la contaminación disminuye.
Igualmente es curioso cómo cíclicamente la contaminación experimenta picos y valles. Estos picos se producen principalmente en invierno, y están bien estudiados, hay pocas lluvias y vientos, existe más producción de electricidad procedente de combustibles fósiles y poca energía de hidroeléctricas y eólica.
En los meses de verano en cambio, la industria suele parar o bajar el nivel de producción, así como la construcción lo que se aprecia también en el dashboard.

6.- Conclusiones

Repetir lo comentado en la introducción, en parte a causa de la insuficiencia de datos, y el corto periodo que se tiene de alguno de ellos, no se ha podido establecer de forma inequívoca una relación directa de que ciertas sustancias químicas contaminantes en la atmósfera estén relacionadas con una mayor o menor tasa de empleo, y por ello con un incremento de las actividades contaminantes más comunes (industria, manufacturero, transporte, construcción…)
Es claro que estos contaminantes sí proceden de estas actividades, pero no ha sido posible probarlo excepto con las partículas PM10, que por otra parte son una muy buena fuente de información de la contaminación atmosférica en todo el país.

Quizá habría que haber estudiado más detenidamente los datos con los que se iba a contar antes de iniciar este proyecto, así como investigar más detenidamente cada uno de los contaminantes, su procedencia y la legislación vigente y futura de emisiones, pues esto ha provocado claramente una información sesgada de resultados.

Por el pequeño número de muestras utilizadas, no ha sido posible aplicar algún otro algoritmo de machine learning más sofisticado, quedándonos en un análisis estadístico básico.

Anexo

Principales fuentes consultadas:

- Instituto Nacional de Estadísticas (INE)
- INE. Definición de ramas de actividad económica
- Tasas de paro por distintos grupos de edad, sexo y comunidad autónoma
- Distribución porcentual de los parados por grupo de edad, sexo y sector económico
- Datos de contaminantes 2001-2016 del MAPAMA
- Normativa europea de la calidad el aire
- Efectos de la polución atmosférica - Organizacion Mundial de la Salud
- Contaminación atmosférica
- Evaluación de la calidad del aire en España 2016
- Datcamp. Visualizing Time Series Data in Python
- Time Series with Plot.ly