INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO
ITAM

Análisis de la Contaminación en la Ciudad de México

Maestría en Ciencia de Datos
Análisis Topológico de Datos

Fernanda Alcalá
Denisse Martínez
José Antonio Ramírez

27 de mayo de 2016

Análisis de la Contaminación en la Ciudad de México

I. ANTECEDENTES

La contaminación en la Ciudad de México ha sido un problema que generó gran atención hace unas décadas dado los efectos palpables en la salud humana y el medio ambiente. A partir de esa coyuntura, el gobierno de la ciudad desarrolló una serie de medidas ambientales, que mostraron año con año un descenso en el índice de calidad del aire. Sin embargo, en los últimos meses se ha agravado al punto que ha replanteado programas ambientales.

Los contaminantes causan daños al sistema nervioso central (partículas, plomo), dolor de cabeza y mareos (CO, SO2), irritación de ojos, nariz y garganta (O3, N02), asma, problemas respiratorios y pulmonares (O3, NO2, PM10, PM2.5), problemas cardiovasculares (partículas, O3), problemas en el hígado, bazo y sangre (plomo, NO2) y problemas de desarrollo fetal (partículas, plomo). Los grupos sensibles a la contaminación son principalmente niños, adultos mayores, deportistas, ciclistas y asmáticos.

El transporte representa el 46% de las contribuciones de emisiones contaminantes, seguido de la industria con un 21%, habitacional con el 20% y el resto con el 13%. Las fuentes contaminantes locales en la CDMX aportan alrededor de la tercera parte de las emisiones que se generan en la Zona Metropolitana del Valle de México.

Los contaminantes principales se clasifican como precursores de ozono, partículas, compuestos tóxicos y compuestos de efecto invernadero.

Contribución de Contaminantes

	Partículas		Precursores de ozono	
	PM10	PM2.5	NOx	COV
Federal y del Estado de México	81%	77%	67%	69%
Distrito Federal	19%	23%	33%	31%

PM10 = Partículas menores a 10 micrómetros
PM2.5 = Partículas menores a 2.5 micrómetros
COV = Compuestos orgánicos volátiles
NOx = Óxidos de nitrógeno

El gobierno de la Ciudad de México cuenta con la Dirección de Monitoreo Atmosférico, que se encarga de recopilar y evaluar información del aire. Los contaminantes principales que se capturan por estación de monitoreo son los siguientes:

a) CO = Monóxido de carbono, medido en ppm (partes por millón)
b) NO = Óxido nítrico, medido en ppb (partes por billón)
c) NO2 = Dióxido de nitrógeno, medido en ppb (partes por billón)
d) NOX = Óxidos de nitrógeno, medidos en ppb (partes por billón)
e) O3 = Ozono, medido en ppb (partes por billón)
f) PM10 = Partículas menores a 10 micrómetros, medidas en µg/m³ (microgramos por metro cúbico)
g) PM2.5 = Partículas menores a 2.5 micrómetros, medidas en µg/m³ (microgramos por metro cúbico)
h) SO2 = Dióxido de azufre, medido en ppb (partes por billón)

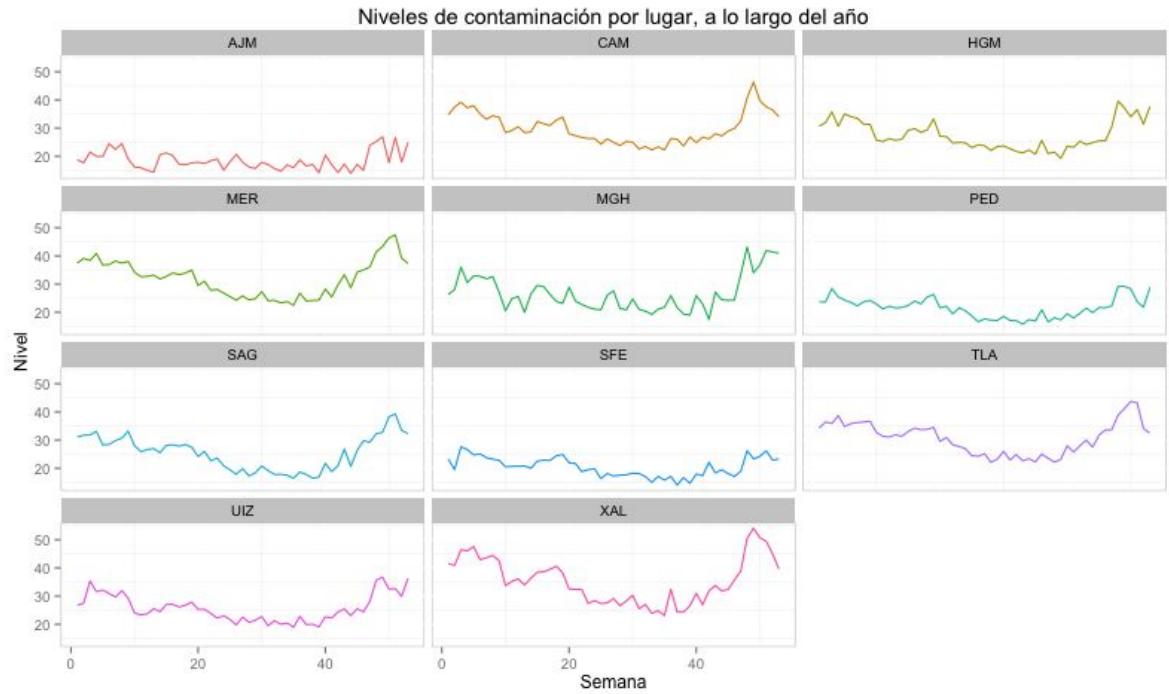
El índice de calidad del aire determinado por las autoridades está compuesto por tres componentes: O3, PM10, PM2.5, por lo que se consideran como variables importantes en el análisis.

OBJETIVO DEL PROYECTO

Realizar análisis topológico de los datos de contaminantes de la Ciudad de México que contribuya a encontrar patrones interesantes y a capturar la dinámica de los datos de forma creativa.

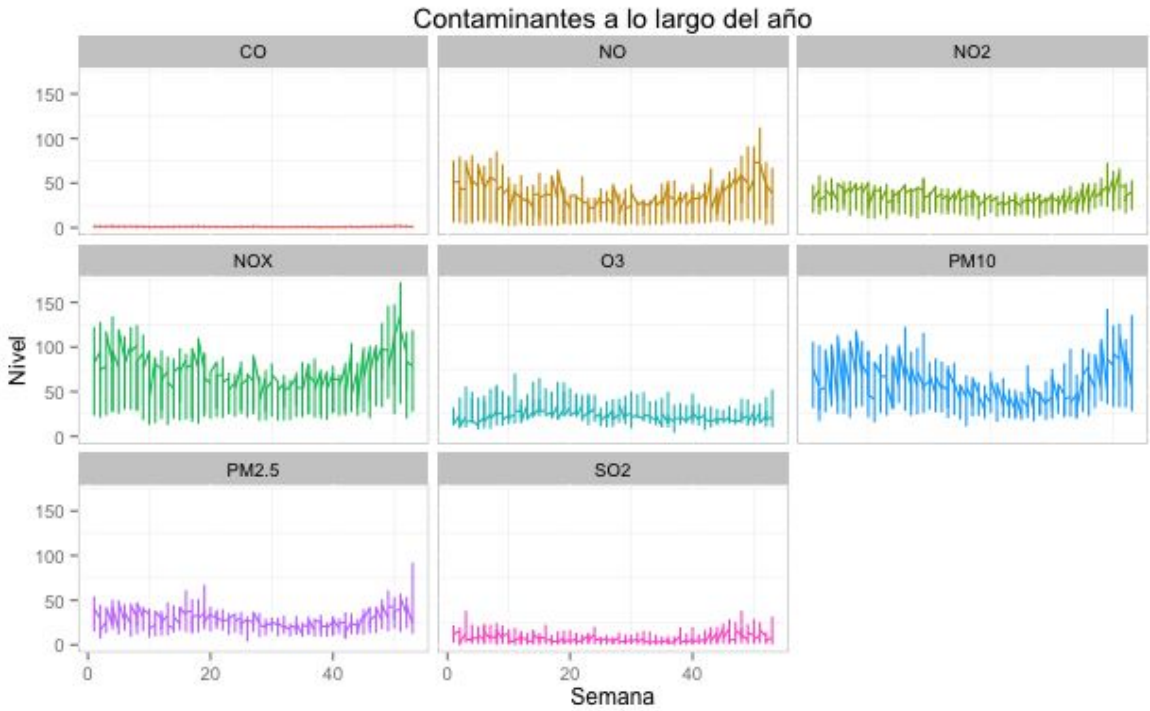
II. ANÁLISIS EXPLORATORIO

Los niveles de contaminación tuvieron un comportamiento similar pero con diferente intensidad según distintas zonas de la ciudad como se puede observar en las siguientes gráficas:



*Las estaciones son Ajusco Medio (AJM), Camarones (CAM), Hospital General (HGM), Merced (MER), Miguel Hidalgo (MGH), Pedregal (PED), San Agustín (SAG), Santa Fe (SFE), Tlalnepantla (TLA), UAM Iztapalapa (UIZ) y Xalostoc (XAL).

Podemos ver una similitud de niveles entre Camarones, Merced, San Agustín y TlaLnepantla. Tanto Merced como San Agustín se encuentran en la zona centro de la ciudad, las otras dos estaciones se encuentran bastante alejadas entre sí. Podríamos sospechar que estas zonas se encontrarán en clusters parecidos. Asimismo, es notorio que Ajusco Medio parece tener los niveles más bajos, lo cual tiene sentido por la ubicación montañosa; por otro lado, Xalostoc tiene los más altos, lo cual es esperado por su cercanía con Tlalnepantla y Ecatepec.



Estas gráficas nos sugieren un comportamiento a lo largo del año que podemos esperar encontrar en los clusters: observamos que los niveles de los contaminantes son consistentemente más altos al final del año, y que comienzan a bajar para llegar a un mínimo alrededor de la semana 30 (finales de julio).

III. METODOLOGÍA

- Análisis exploratorio de datos disponibles y alcance del proyecto

La información de los contaminantes se descargó de la Dirección de Monitoreo Ambiental. La información es recopilada por las 45 estaciones de monitoreo ubicadas en el Valle de México, que capturan información cada hora desde 1986.

- Limpieza, creación y selección de atributos

Una vez identificadas las variables se preprocesaron los datos, limpiando e integrando la información. Se seleccionó información desde 2007 hasta la disponible a 2016, sin embargo los datos disponibles contienen muchos valores NA, por lo que se promedió el nivel de los contaminantes por semana, por estación, generando un total de 2,648 observaciones.

- Análisis de atributos

Con el fin de entender la dinámica de las variables y reducir la dimensionalidad, se efectúa un Análisis de Componentes Principales.

- Definición del algoritmo de agrupamiento

Se seleccionó trabajar con el algoritmo Vecinos Más Cercanos que tiene como fin clasificar cada dato según el grupo más común al que pertenecen sus vecinos.

Una parte fundamental del análisis consistió en emplear la componente 1 para dividir la información en intervalos. Se aplicó el algoritmo de agrupamiento a cada intervalo y posteriormente cada intervalo generó clusters.

- Caracterización de grupos

Una vez identificados los elementos que pertenecen a cada grupo, se analizan sus características.

- Visualización e interpretación de resultados

Para facilitar la interpretación se presenta la visualización de TDA_MAPPER.

ANÁLISIS DE COMPONENTES PRINCIPALES

Se realizó un Análisis de Componentes Principales (PCA por sus siglas en inglés) para entender cómo es que diferentes variables explican la dinámica de la información. Además de reducir la dimensionalidad y la redundancia en los datos.

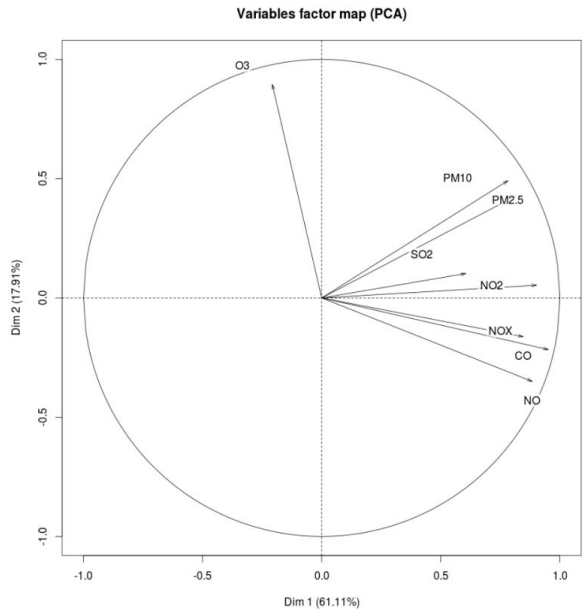
Consideraciones de PCA:

- Sólo encuentra relaciones lineales.
- Asume que la covarianza es un buen indicador de las relaciones interesantes entre atributos. Por lo cual si los datos no pueden ser adecuadamente descritos a través de su media y su varianza el método es inadecuado (si las observaciones no se distribuyen de forma gaussiana o exponencial).
- Componentes ortogonales.

En general, el método PCA consiste en encontrar una matriz P de tal forma que la matriz de covarianza PX es diagonal y las entradas de esa diagonal tengan orden descendente. PX representará las variables transformadas que son combinaciones lineales de los atributos originales y estarán conformadas por los pesos calculados en P. Dado que la matriz de covarianza es diagonal, la covarianza de cualquier par de variables distintas será cero, por lo que las varianzas de las nuevas variables estarán sobre la diagonal.

Los resultados que arrojó el PCA para los 8 atributos, indican que el 79.02% de la varianza lo explican los primeros dos componentes.

Variable	PC1	PC2
CO	0.8458752	-0.16262478
NO	0.8836539	-0.34995749
NO2	0.9016207	0.05347668
NOX	0.9509814	-0.2166083
O3	**-0.20745**	0.89519052
PM10	0.8096371	0.42491088
PM2.5	0.7816732	0.49187542
SO2	0.6044153	0.1022048



Analizando la tabla anterior y la gráfica anterior se puede mostrar que la primera componente muestra una correlación positiva con todos los contaminantes a excepción de la variable O3 (Ozono), que tiene una correlación negativa.

La primer componente será empleada como variable para dividir la información en grupos antes de aplicar la técnica de agrupamiento.

ALGORITMO “VECINOS MÁS CERCANOS” IMPLEMENTADO CON BREADTH-FIRST SEARCH

El algoritmo de vecinos más cercanos agrupa observaciones que están a cierto grado de otras observaciones cercanas. Para este análisis se empleó una versión del algoritmo desarrollada por el Profesor Mauricio García Tec, el cual realiza una búsqueda *breadth-first* para explorar la estructura del grafo y sus uniones. Este algoritmo reduce considerablemente la complejidad de la exploración de datos, pues al utilizar una *cola* (queue), itera solamente una vez en el conjunto de nodos.

A grandes rasgos, el algoritmo de búsqueda *breadth-first* comienza en un nodo, y asigna una distancia infinita a todos los demás. Después, encuentra el nodo más cercano entre los nodos conectados a él y lo marca como visitado. De esta manera, va explorando el grafo por medio de iteración en todos los nodos no visitados. Es un algoritmo ampliamente usado para encontrar los caminos más cercanos entre pares de nodos

Más adelante, este algoritmo de agrupación ayudará a no necesariamente añadir una observación a un grupo, aún y cuando se encuentre en el rango de “vecino” sino que dependerá además de la distancia para agruparlo. Esto permite identificar clusters de una manera más precisa, pues al no obligar a tener un número fijo de observaciones, permite conservar estructuras de clusters más pequeños o menos dispersos que otros.

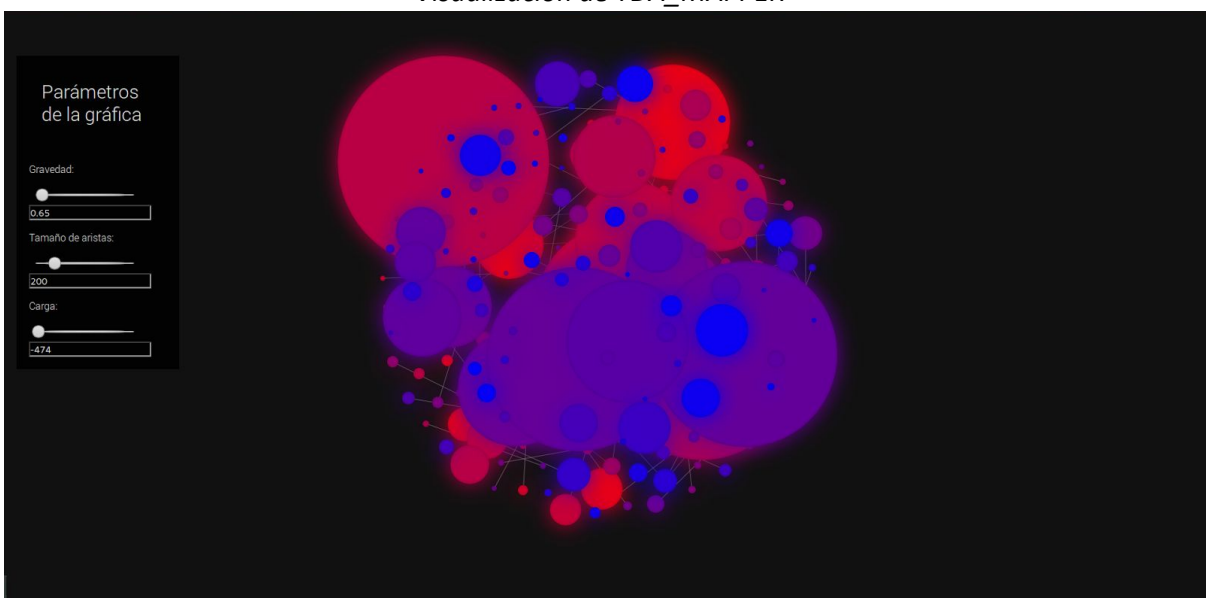
V. RESULTADOS

El algoritmo de agrupamiento generó 198 clusters. Este es un número considerablemente alto como para analizar detalladamente, así que es un excelente problema para atacar con algunas técnicas de Análisis Topológico de Datos que se vieron durante el semestre.

Después de intentar con todas las variables, hemos definido la escala de color de la gráfica por la variable PM2.5_prom, la cual se refiere a contaminación por partículas pequeñas (2.5 micrómetros de diámetro).

Obtuvimos una coloración interesante. Hace incluso evidente a primera vista que existen solamente dos clusters "gigantes" que agrupan a los 198 clusters originales.

Visualización de TDA_MAPPER



Analizando en general los clusters, se observa que los grupos de color azul representan un nivel bajo de PM2.5 y además agrupan a semanas de principios y finales de año, mientras que los clusters en color rojo agrupan aquellas semanas donde los contaminantes son altos. Esto es consistente con observaciones empíricas: los niveles de contaminación son mayores en los meses cálidos.

VI. REFERENCIAS

Dirección de Monitoreo Ambiental, *Contaminante,* consultado el 22/05/2016 en http://www.aire.df.gob.mx/default.php?opc=%27aKBhnmI=%27&opcion=Zg==

United States Environmental Protection Agency, *Basic Information,* consultado el 26/05/2016 en https://www3.epa.gov/pm/designations/basicinfo.htm