Instituto Tecnológico y de Estudios Superiores de Monterrey **MA2007B**

Grupo 602

Examen Práctico: Problema 3

Contaminación del Aire en Nueva York

Docente:

Manuel Alejandro Ucán Puc

Autoría:

David Fernando Armendáriz Torres
A01570813

Introducción

La problemática que se aborda a lo largo de este escrito corresponde al tema de la calidad del aire. Se busca estudiar la calidad del aire en una ciudad y determinar si existe una relación entre la geografía de la ciudad y la calidad del aire.

Dadas las limitaciones de tiempo presentes, se optó por acotar la problemática y se planteó la siguiente pregunta de investigación:

¿Existe alguna relación entre la calidad del aire con base en contaminantes selectos (NO_2, O_3, PM 2.5) y la ubicación geográfica de las subdivisiones de la ciudad de Nueva York (UHF42, CD)?

El análisis correspondiente se llevó a cabo sobre una base de datos con registros referentes a la calidad del aire en la ciudad de Nueva York *Air Pollution Dataset (Maharaj, 2024).* Sobre estos datos se aplicó un algoritmo Mapper utilizando la librería *kmapper* de Python con la intención de observar propiedades topológicas de los datos al generar complejos simpliciales con base en la ubicación geográfica de los registros.

Preprocesamiento de Datos

La base de datos empleada cuenta con un total de 12 variables y 16,218 registros. Como parte del preprocesamiento de los datos, se depuró eliminando múltiples columnas y filas irrelevantes para el análisis. A continuación, se describen aquellos datos que se conservaron:

Variables:

- Unique ID: identificador único
- Name: indicador categórico del tipo de dato registrado
- Geo Join ID: identificador categórico del área geográfica.
- Geo Type Name: identificador categórico del tipo estándar geográfico empleado por Geo Join ID

Registros:

- Name correspondiente a “Nitrogen dioxide (NO_2)", "Fine particles (PM 2.5)" y "Ozone (O_3)”.
- Geo Type Name correspondiente a "UHF42", "UHF34" y "CD".

Tras esta filtración se conservan 12,420 registros, 5,265 de estos son NO_2, otros 5,265 PM 2.5 y los restantes 1,890, O_3.

Adicionalmente, el estándar UHF34 se transformó a UHF42 segmentando las subdivisiones conjuntas (105106107, 404406, 501503, 306308, 309310, 305307, 503504, 501502) en uno de sus elementos constituyentes seleccionados de manera arbitraria (106, 405, 502, 307, 309, 306, 504, 502).

Dadas las diferencias en los estándares UHF42 y CD, se optó por segmentar los datos con base en dichos estándares y emplear el análisis sobre cada segmento para no introducir ruido innecesario en el modelado.

De manera similar, para realizarse un análisis de mayor precisión bajo los contaminantes seleccionados, se segmentaron los datos con base en el contaminante y se empleó el análisis sobre cada contaminante.

Finalmente, los datos, originalmente dataframes de *pandas*, se convirtieron a arreglos de *numpy* acorde a los requisitos del algoritmo Mapper.

El resultado final fueron 6 arreglos de numpy a analizar mediante el modelado posterior:

- NO2 UHF42
- NO2 CD
- PM 2.5 UHF42
- PM 2.5 CD
- O3 UHF42
- O3 CD

Modelado

Para el modelado, como se mencionó previamente, se utilizó un algoritmo Mapper utilizando las librerías *kmapper*, *numpy* y *sklearn* de Python.

El algoritmo empleado se generó con los siguientes hiperparámetros:

- Proyección: Geo Join ID
- Cubierta:
  - Cantidad de cubos: 5
  - Overlap: 0.7
- Mapper:
  - Agrupador: K-Means
    - Grupos: 8

Los hiperperámetros empleados se seleccionaron mediante prueba y error buscando el mejor desempeño del método.

Por cada arreglo empleado, se generaron dos visualizaciones una con coloración a base de la concentración del contaminante en cuestión y otra con coloración a base del *Borough* al cual pertenecen los registros geográficos.

Los *Boroughs* corresponden a El Bronx(Rojo) de orden 100, Queens(Verde) de orden 400, Staten Island(Amarillo) de orden 500, Brooklyn(Morado) de orden 200 y Manhattan(Azul) de orden 300.

Análisis de Resultados

El modelado planteado generó complejos correspondientes a los 5 *Boroughs* empleados lo cual permitió el análisis de concentración de contaminantes.

En el análisis consiguiente se adjuntan solo 2 figuras, correspondientes al modelado de NO_2 bajo el estándar CD. El resto de los mapeos se adjuntan como anexos al final del reporte dentro de un repositorio con todos los archivos pertinentes.

NO_2, CD

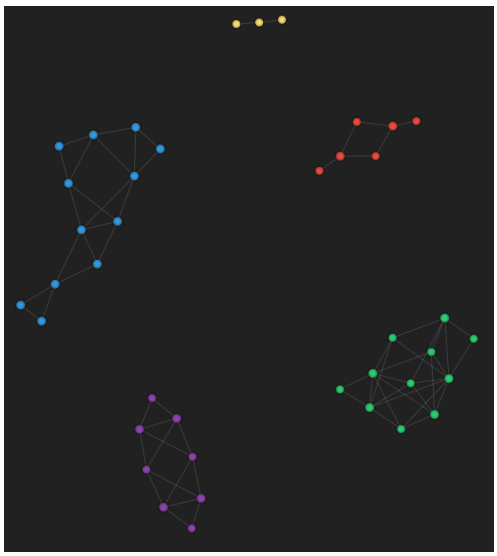

Ilustración 1 Mapeo sobre Geo ID y coloración a base del Borough.

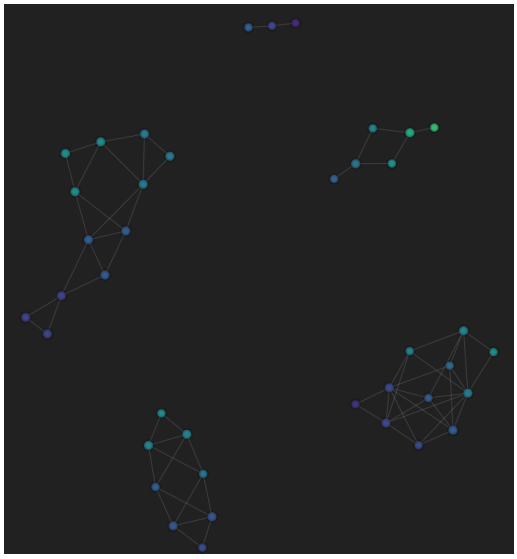

Ilustración 2 Mapeo sobre Geo ID y coloración a base de concentración de NO2. A mayor claridad, mayor concentración.

Del mapeo se observa una alta concentración de NO_2 en la región de Bronx, donde a su vez se encuentran los registros máximos. Por contraparte, la menor concentración se presenta en Staten Island, mientras que el resto de las regiones no presentan una decantación significativa, es decir una concentración moderada.

Contaminantes Restantes

Con la brevedad en mente, el resto de los resultados se presentan a modo de tabla.

CD	PM 2.5	O_3
Alta concn.	-	Queens St. Island Manhattan Brooklyn
Baja concn.	Queens St. Island Manhattan Brooklyn	-
Máxima concn.	Bronx	Queens
Moderada concn.	Bronx	Bronx

Tabla 1 Resultados de mapeo por estándar CD. concn: concentración.

UHF42	NO_2	PM 2.5	O_3
Alta concn.	Bronx Manhattan	Manhattan	Bronx Brooklyn Queens St. Islan
Baja concn.	-	St. Island Queens	-
Max concn.	Bronx	Manhattan	Queens
Mode-rada concn.	St. Island Queens Brooklyn	Bronx Brooklyn	Manhattan

Tabla 2 Resultados de mapeo por estándar UHF42. concn: concentración.

Conclusiones

Con base en los resultados planteados se puede afirmar que la ubicación geográfica efectivamente está relacionada con la contaminación del aire en Nueva York. Es decir algunas regiones de la ciudad

cuentan con mayore concentración de contaminantes. No obstante, la influencia de la región no parece ser muy extrema, simplemente parece denotar una leve tendencia hacia altas o bajas concentraciones en algunos casos. En términos mas generales, la región de Bronx, Brooklyn y Manhattan tienden un poco a una concentración de contaminantes media alta, Queen a una concentración media y Staten Island a una concentración media baja.

Adicionalmente, parece ser que el incremento de concentración dentro de los complejos se debe en parte a la presencia de los valores máximos de los datos.

Por último, las limitaciones presentes en el análisis generan muchas áreas de oportunidad para futuras investigaciones. Sería de utilidad, por ejemplo:

- Transformar el estándar CD a UHF42 o viceversa para analizar todos los datos en conjuntos.
- Manejar mismas unidades y escalas en los 3 contaminantes para poder comparar sus concentraciones.

Referencias

Maharaj, S. (2024, mayo). *Air Pollution Dataset*. Kaggle. https://www.kaggle.com/datasets/sahirmaharajj/air-pollution-dataset/data

Anexos

Repositorio GitHub (Archivos):

https://github.com/David-A-T/MA2007B.602-Examen-Practico