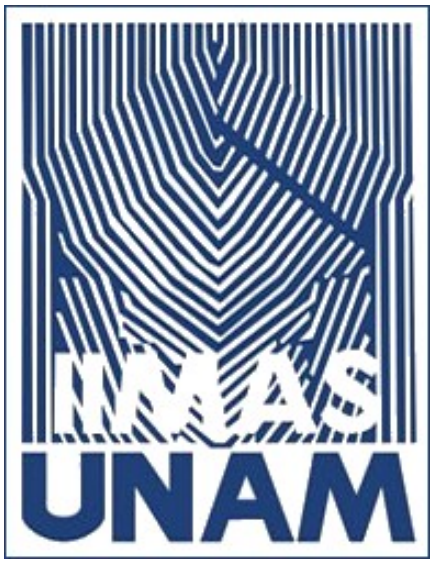
Universidad Nacional Autónoma de México
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Licenciatura en Ciencia de Datos



Análisis Discreto en la contaminación del aire en la Ciudad de México

Raúl Isaid Mosqueda García

Matemáticas Discretas

Enero 2020
Ciudad de México, México

RESUMEN

En este proyecto se realizo un análisis mediante la implementación de algoritmos discretos a datos recabados por el Gobierno de la Ciudad de México por un conjunto de estaciones distribuidas dentro del Valle de México. Los datos corresponden a los indicies de los contaminantes en el aire de Ozono(O_3), Dióxido de Azufre(SO_2), Dióxido de Nitrógeno(NO_2), Monóxido de Carbono (CO), y la suspensión de partículas PM10, recabadas para los años correspondientes desde el 2000 hasta el 2018.

Cada uno de estos contaminantes representan amenazas contra la salud de los ciudadanos que puedan intoxicarse si existe una alta concentración de alguno de estos en el aire que se respira, donde los problemas de salud van desde una reacción alérgica hasta desarrollos de problemas asmáticos, contaminación del oxigeno en la sangre hasta reacciones letales para concentraciones muy severas.

El texto viene segmentado en 2 partes, la primera consiste de un análisis elemental donde cualquier persona puede en general, entender los conceptos, metodología y resultados presentados, mientras que en la segunda se hace un análisis formal y riguroso sobre los mismos puntos, donde se requiere un conocimiento previo sobre desarrollo y análisis de algoritmos, así como ser cómodos con la formalidad matemática y ligeros tecnicismos químicos.

El texto esta escrito de forma que mayormente ambas secciones puedan ser leídas de manera individual, por esto algunas secciones parecen redundantes en su mayoría, sin embargo cualquiera de los dos es suficiente por si mismo para comprender el análisis realizado.

Palabras Clave **–Ozono, Dióxido de Azufre, Dióxido de Nitrógeno, Monóxido de Carbono, partículas por billón, partículas por millón**.

NOMENCLATURA

- ppb: partículas por billón.
- ppm: partículas por millón.
- $\mu g/m^3$: micro-gramos por metro cubico.
- lis: Secuencia creciente mas larga (longest increasing sequence).

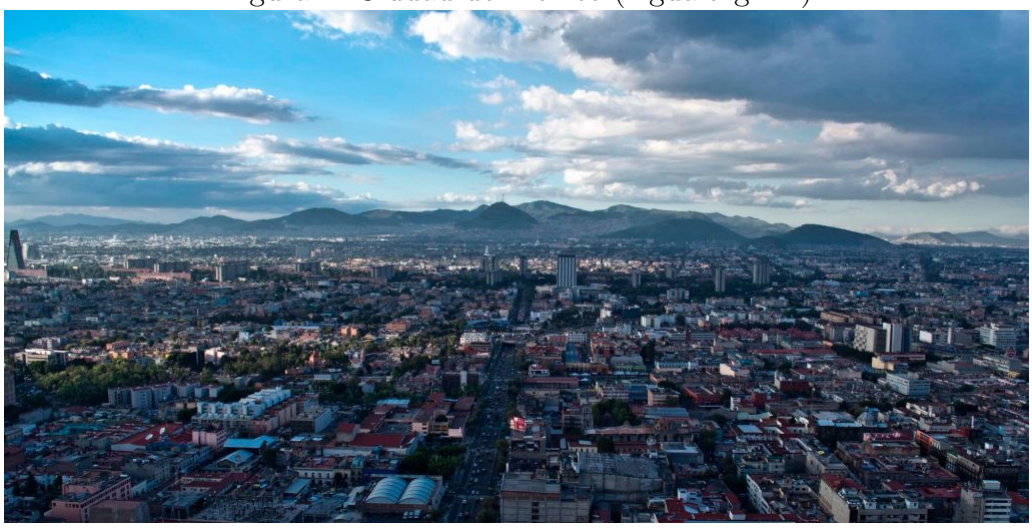
Índice general

Documentación Ejecutiva

0.1. Problema a resolver

El Valle de México es una región compuesta por la Ciudad de México y gran parte del Estado de México. Dentro del Valle de México habitan aproximadamente $19,768,740$ personas (INEGI,2013), las cuales todas comparten una ”atmósfera” exclusiva para la población dentro del valle.

Figura 1: Ciudad de México (Agua.org.mx)



La atmósfera particular del Valle de México se ocasiona debido a su elevada altitud y a que se encuentra rodeado por montañas lo cual ocasiona que los gases y compuestos que se encuentran mezclados con el aire del Valle, se mantengan dentro por una cantidad considerable de tiempo, es decir, que el aire dentro del valle se mantiene por largos periodos en el valle y no es fácil que este se limpie con el aire de las

regiones vecinas, pues toma tiempo llegar hasta dentro del valle.

Por otro lado, debido a la rápida urbanización de la ciudad de México desde los años 70's y la alta proporción de inmigración hacia el valle ha logrado que el Valle de México sea una región extremadamente densa en población, que junto con el crecimiento de la industrialización ha hecho del valle de México una región importante en lo que respecta al cuidado de su ecosistema, pues este ha sido desplazado drásticamente durante las ultimas décadas.

La alta densidad de población junto con la implementación de métodos de transporte y procesamiento de materiales que emiten al aire compuestos nocivos para la salud, han puesto ya en repetidas ocasiones en estados de alerta a la población del valle puesto que la calidad del Aire es tan pobre que empieza a ser nociva para algunas secciones de la población. Entre los principales componentes nocivos encontrados en el aire que se monitorizan constantemente (cada hora) dentro del valle de México son:

1. **Ozono** (O_3). El Ozono es un compuesto que se forma por varias reacciones derivadas principalmente de las emisiones de gases vehiculares e industriales donde la gasolina y solventes orgánicos forman parte del proceso.

   La exposición continua a altas concentraciones de ozono puede causar daños permanentes en los pulmones. En individuos sensibles, la exposición a niveles bajos de ozono puede ocasionar tos, náusea, irritación en las mucosas de nariz y garganta y congestión en vías respiratorias. Asimismo, puede exacerbar problemas pre-existentes de salud como bronquitis, enfisema, asma y enfermedades cardiacas. La población más sensible a estos problemas son frecuentemente los niños, los ancianos y personas que realizan actividades al aire libre por largos períodos.

2. **Dióxido de Azufre** (SO_2).El dióxido de azufre es la combustión de productos petrolíferos y la quema de carbón en centrales eléctricas y calefacciones centrales. Existen también algunas fuentes naturales, como es el caso de los volcanes.

la exposición a altas concentraciones de dióxido de azufre durante 5 a 10 minutos en pacientes asmáticos sometidos a condiciones de ventilación incrementadas, a partir de las 200 ppb, ocasionan un aumento en la presencia de síntomas respiratorios (sibilancias, tos, dificultad para respirar) y una disminución en la función pulmonar, además de un aumento en marcadores de inflamación a nivel. Asimismo, se ha observado una respuesta broncoconstrictora derivada de la exposición a dióxido de azufre en pacientes con asma, la cual inicia desde los primeros minutos de exposición y tiende a disminuir posterior a 1 hora si se reduce dicha exposición.

3. **Dióxido de Nitrógeno** (NO_2). El origen del dióxido de nitrógeno es en general, el residuo de procesos de combustión.

   La acumulación de dióxido de nitrógeno en el cuerpo humano, constituye un riesgo para las vías respiratorias ya que se ha comprobado que: inicia, reactiva y puede alterar la capacidad de respuesta de las células en el proceso inflamatorio, como sucede con las células polimorfonucleares, macrófagos alveolares y los linfocitos, siendo más frecuente en casos de bronquitis crónica.

4. **Monóxido de Carbono** (CO). El monóxido de carbono similarmente al dióxido de nitrógeno es el residuo de procesos de combustión, sin embargo el monóxido de carbono deriva directamente de de la combustión incompleta de compuestos de carbono, en el Valle de México su principal fuente es la combustión de los vehículos automotores así como la industria.

   El riesgo de la exposición al monóxido de carbono varía desde el efecto de pequeñas cantidades atmosféricas en individuos que padecen deficiencias circulatorias (siendo particularmente susceptibles los enfermos con angina de pecho, así como aquellos con arterioesclerosis), hasta una intoxicación aguda por inhalación de grandes cantidades del contaminante en espacios cerrados y/o en un período corto.

5. **Particulas PM10**. Las partículas se definen como cualquier material que existe

en estado sólido o líquido en la atmósfera o en una corriente de gas, excepto agua o hielo. Las partículas incluyen polvo, ceniza, hollín, humo y pequeñas partículas de contaminantes. De acuerdo con su diámetro aerodinámico, se clasifican en menores o iguales a 10 µm (PM10). Contienen principalmente materiales de la corteza terrestre y se originan en su mayoría por procesos de desintegración de partículas más grandes. También pueden contener material biológico como polen, esporas, virus o bacterias o provenir de la combustión incompleta de combustibles fósiles.

Toda la información mencionada anteriormente fue citada directamente de la pagina de la SEMARNAT y Instituto para la Salud Geoambiental: `http://dgeiawf.semarnat.gob.mx:8080/ibi_apps/WFServlet?IBIF_ex=D3_R_AIRE01_01&IBIC_user=dgeia_mce&IBIC_pass=dgeia_mce`, `https://www.saludgeoambiental.org/`.

Con esta explicación recabada, lo que se busca entonces es entonces realizar un análisis de la información existente respecto a las cantidades de estos componentes nocivos para la salud, y en general encontrar si han existido situaciones criticas en el pasado y cuales han sido las acumulaciones mas largas de estos componentes . Con esto en mente, se procede a la elección de la base de datos.

0.2. Elección de la base de datos

Para este análisis se decidió trabajar con la base de datos recabada por el gobierno de la ciudad de México, la base es

”Indice de Calidad del Aire (horarios)” y se puede consultar en el siguiente enlace: `http://www.aire.cdmx.gob.mx/default.php?opc=%27aKBhnmI=%27&opcion=aw==`.

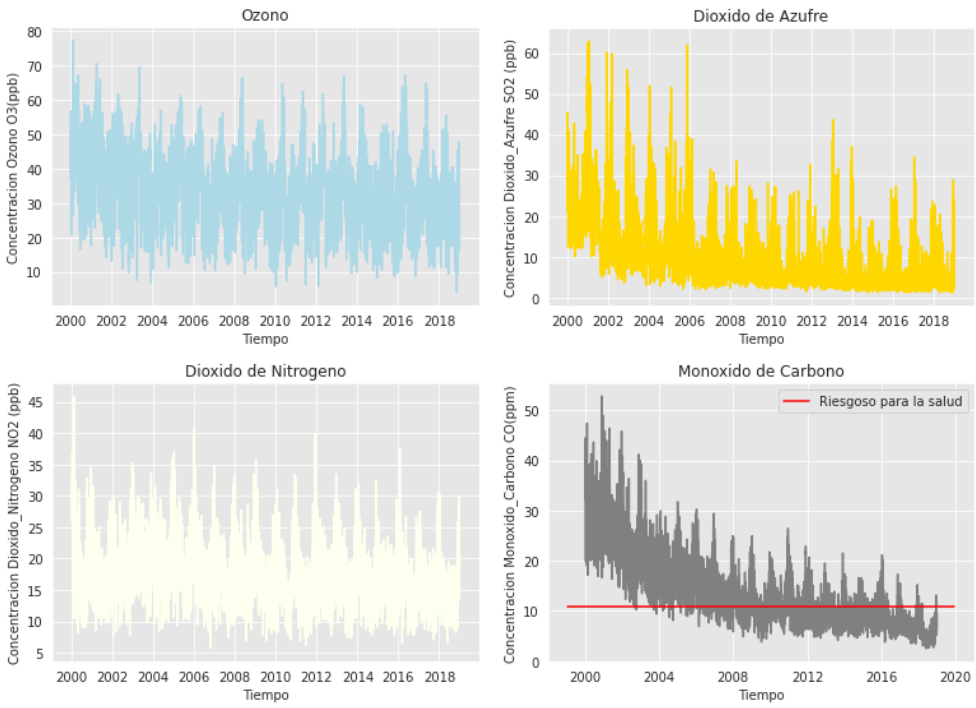
La base contiene los índices de los componentes nocivos recabados por estaciones atmosféricas distribuidas alrededor del Valle de México, estos índices son calculados de manera diaria promediadas por hora las 24 horas del día, desde 1990 hasta el año 2018.

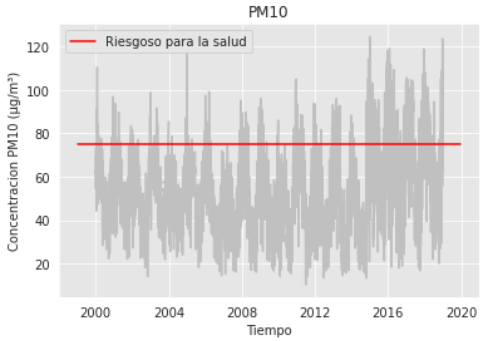
La base es apropiada puesto que contiene información bien estructurada para realizar el análisis que se requiere, además de que los datos son obtenidos directamente de un conjunto de estaciones, la recompilación de los datos se realiza físicamente de las mediciones obtenidas para conjuntos de aire que posteriormente son promediadas entre un grupo de estaciones, por lo que difícilmente se presta a obtener variaciones anormales o ocasionadas por alguna clase de ruido.

Para este análisis se trabajara con los datos correspondientes a los años posteriores (e incluyendo) a los del año 2000, posteriormente estos serán manejados de la forma que se requiera para el análisis.

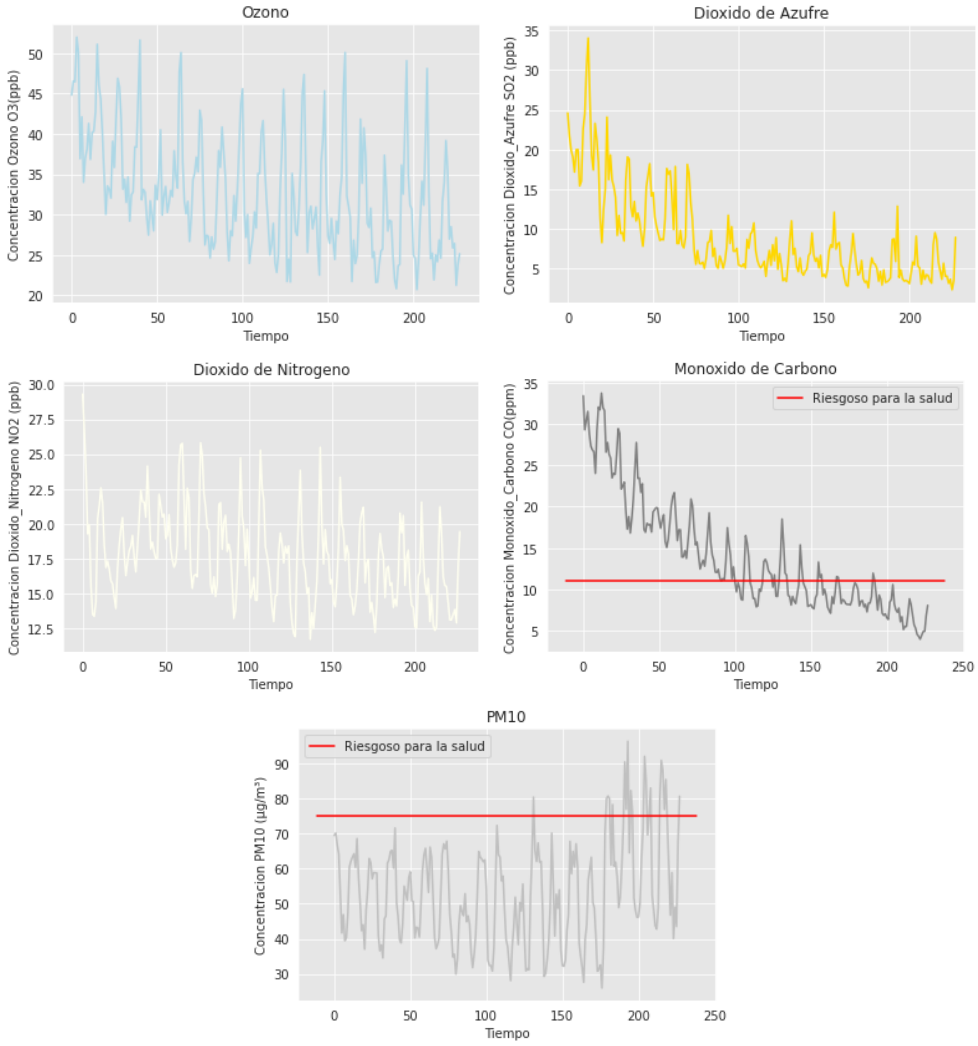
0.3. Metodología

Lo primero que se realizo fue la reducción de los datos a promedios por mes puesto que los promedios por hora implican demasiada información para procesar. Los datos diarios se ven de la siguiente forma:





los cuales se ven de manera mas clara cuando se promedian por mes:



Con los datos obtenidos en meses se paso a realizar los análisis descritos previamente.

0.3.0.1. Acumulaciones mas largas

Una forma de comprender la gravedad de la circunstancia en la que se encuentra la población del Valle de México, es encontrar la cantidad de días que han tomado los incrementos mas grandes de la concentración de los componentes, estos incrementos se pueden ver como en que momentos se tuvieron días donde para en alguno de los próximos se obtuvo una concentración mayor, hasta encontrar todos los días donde esto ocurre.

De forma general hay 2 formas de obtener estas acumulaciones mas largas. La primera consiste en ir registrando una acumulación mas grande y conforme se vayan procesando los datos puede que surjan acumulaciones mas largas que la considerada anteriormente, entonces la nueva se considera como la acumulación mas grande, así hasta que se hayan analizado todos los datos. Este método en general resulta ser muy intuitivo pero muy complicado de realizar cuando se tienen cantidades muy grandes de datos, por lo que para este análisis se realizo mediante la segunda forma, sin embargo esta sera explicada en la documentación técnica ya que este la explicación va mas allá de los objetivos de esta sección.

Las concentraciones son entonces las siguientes:

- **Ozono** (O_3).($min = 24.63, max = 33.959307795698926, meses\_crecientes : 8$) Estos rangos indican que si bien se mantuvo un tiempo de al menos 8 meses dentro de un crecimiento constante de concentración, se mantuvo por debajo de la mitad del rango mas grande, lo cual en sentido de la acumulación mas larga, indica que se tuvo un buen comportamiento históricamente.

- **Dióxido de Azufre** (SO_2).($min = 2.85, max = 8.9, meses\_crecientes : 3$) Para este componente se tuvo un excelente comportamiento pues la concentración mas grande fue de mínimo 3 meses, en un rango bastante bajo comparado con el rango mas grande registrado.

- **Dióxido de Nitrógeno** (NO_2).($min = 11.7, max = 21.21, meses\_crecientes : 14$)

En lo que respecta al Nitrógeno se tuvo una mal comportamiento pues se mantuvo por al menos 14 meses en un rango demasiado grande respecto al rango total, afortunadamente este componente en general nunca alcanza niveles críticos para la salud.

- **Monóxido de Carbono** (CO).($min = 3.98, max = 8.07, meses\_crecientes : 6$) Al igual que los casos anteriores, en respecto a la acumulación mas grande, esta además se mantuvo totalmente por debajo de la sección critica, sin embargo estuvo cerca de alcanzarla, por lo que respecto a la acumulación mas larga se mantuvo un comportamiento saludable.

- **Partículas PM10**.($min = 25.98, max = 80.64, meses\_crecientes : 10$) El PM10 tuvo un comportamiento extremadamente peligroso puesto que se mantuvo por al menos 10 meses en un intervalo bastante grande el cual rebasa la marca de la sección nociva para la salud. Este resultado es alarmante para la población puesto que además las partículas PM10 resultan ser de los contaminantes mas letales para la mayoría de las personas.

Si se desean replicar estos resultados, resulta bastante fácil puesto que a pesar demasiados datos estos se encuentran en un formato el cual no toma demasiado espacio de memoria para almacenarlos. Análogamente, el procesamiento de los datos no requiere mucha capacidad de procesamiento puesto que aun no representan una cantidad significativa donde se tengan que realizar operaciones elementales (suma y división) de forma que se requiera un poder sofisticado de computación.

Para realizar puntualmente el análisis puede que si se requiera una capacidad considerable de memoria inmediata (RAM) y poder de procesamiento arriba del promedio para la primera forma de realizar el análisis pues sus requerimientos crecen demasiado respecto a la cantidad de datos que se vayan a procesar, sin embargo para la implementación que se realizo en este análisis (la cual se describe a detalle en la próxima sección) estos problemas se resuelven puesto que de esa forma no se incrementan tanto los requerimientos a diferencia de como ocurre con la implementación intuitiva.

0.4. Conclusiones

Este análisis permitió obtener resultados que no resultan tan intuitivos respecto a los niveles de contaminantes en el aire. Obtener las acumulaciones mas grandes permiten obtener una idea de lo que son los tamaños de periodos de tiempos donde en general, se tiene la tendencia de que cada día siguiente sera peor que el anterior, obtener los intervalos y la longitud de estas concentraciones dan una explicación de cuales son los peores casos constantes y que es lo peor que se podría esperar.

Para este análisis se vio entonces que entre los principales contaminantes nocivos para los humanos las peores situaciones son relativamente aceptables para la mayoría de los contaminantes, sin embargo para los 2 mas letales que son el Dióxido de Nitrógeno y las partículas PM10 se han tenido casos extremadamente peligrosos, por lo que se espera que se hagan los cambios necesarios en los estilos de vida de las personas para mejorar la prevención de la emisión de estos contaminantes, así como la prevención de su consumo mediante el aire.

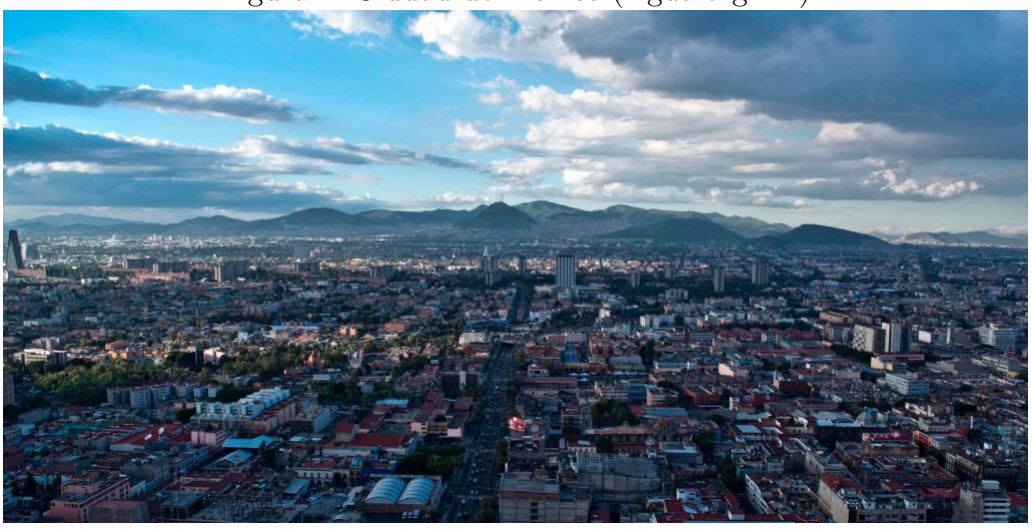
Idealmente se espera que las personas que hayan realizado la lectura de este proyecto se vuelvan mas conscientes en respecto a que actividades realizan que pueden producir estos contaminantes y como evitarlos, puesto que se vio que son extremadamente peligrosos para la salud e históricamente ya se han tenido situaciones criticas, por la salud propia y de los demás se motiva a hacer conciencia sobre el tema tratado.

Documentación Técnica

0.5. Problema a resolver

El Valle de México es una región compuesta por la Ciudad de México y gran parte del Estado de México. Dentro del Valle de México habitan aproximadamente $19,768,740$ personas (INEGI,2013), las cuales todas comparten una ”atmósfera” exclusiva para la población dentro del valle.

Figura 2: Ciudad de México (Agua.org.mx)



La atmósfera particular del Valle de México se ocasiona debido a su elevada altitud y a que se encuentra rodeado por montañas lo cual ocasiona que los gases y compuestos que se encuentran mezclados con el aire del Valle, se mantengan dentro por una cantidad considerable de tiempo, es decir, que el aire dentro del valle se mantiene por largos periodos en el valle y no es fácil que este se limpie con el aire de las

regiones vecinas, pues toma tiempo llegar hasta dentro del valle.

Por otro lado, debido a la rápida urbanización de la ciudad de México desde los años 70's y la alta proporción de inmigración hacia el valle ha logrado que el Valle de México sea una región extremadamente densa en población, que junto con el crecimiento de la industrialización ha hecho del valle de México una región importante en lo que respecta al cuidado de su ecosistema, pues este ha sido desplazado drásticamente durante las ultimas décadas.

La alta densidad de población junto con la implementación de métodos de transporte y procesamiento de materiales que emiten al aire compuestos nocivos para la salud, han puesto ya en repetidas ocasiones en estados de alerta a la población del valle puesto que la calidad del Aire es tan pobre que empieza a ser nociva para algunas secciones de la población. Entre los principales componentes nocivos encontrados en el aire que se monitorizan constantemente (cada hora) dentro del valle de México son:

1. **Ozono** (O_3). El Ozono es un compuesto que se forma por varias reacciones derivadas principalmente de las emisiones de gases vehiculares e industriales donde la gasolina y solventes orgánicos forman parte del proceso.

   La exposición continua a altas concentraciones de ozono puede causar daños permanentes en los pulmones. En individuos sensibles, la exposición a niveles bajos de ozono puede ocasionar tos, náusea, irritación en las mucosas de nariz y garganta y congestión en vías respiratorias. Asimismo, puede exacerbar problemas pre-existentes de salud como bronquitis, enfisema, asma y enfermedades cardiacas. La población más sensible a estos problemas son frecuentemente los niños, los ancianos y personas que realizan actividades al aire libre por largos períodos.

   Los límites vigentes son: una concentración menor o igual a 0.095 ppm considerando el máximo horario y 0.070 ppm considerando el máximo de los promedios móviles de 8 horas, en un año.

2. **Dióxido de Azufre** (SO_2).El dióxido de azufre es la combustión de productos petrolíferos y la quema de carbón en centrales eléctricas y calefacciones centrales. Existen también algunas fuentes naturales, como es el caso de los volcanes.

   la exposición a altas concentraciones de dióxido de azufre durante 5 a 10 minutos en pacientes asmáticos sometidos a condiciones de ventilación incrementadas, a partir de las 200 ppb, ocasionan un aumento en la presencia de síntomas respiratorios (sibilancias, tos, dificultad para respirar) y una disminución en la función pulmonar, además de un aumento en marcadores de inflamación a nivel. Asimismo, se ha observado una respuesta broncoconstrictora derivada de la exposición a dióxido de azufre en pacientes con asma, la cual inicia desde los primeros minutos de exposición y tiende a disminuir posterior a 1 hora si se reduce dicha exposición.

   Los límites normados máximos de concentración del dióxido de azufre para proteger la salud pública son: 0.075 ppm para el límite de 1 hora considerado como el promedio aritmético de 3 años consecutivos de los percentiles 99 anuales, obtenidos de los máximos diarios; 0.04 ppm para el límite de 24 horas considerado como el máximo de 3 años consecutivos.

3. **Dióxido de Nitrógeno** (NO_2). El origen del dióxido de nitrógeno es en general, el residuo de procesos de combustión.

   La acumulación de dióxido de nitrógeno en el cuerpo humano, constituye un riesgo para las vías respiratorias ya que se ha comprobado que: inicia, reactiva y puede alterar la capacidad de respuesta de las células en el proceso inflamatorio, como sucede con las células polimorfonucleares, macrófagos alveolares y los linfocitos, siendo más frecuente en casos de bronquitis crónica.

   La concentración de dióxido de nitrógeno, como contaminante atmosférico, no debe rebasar el límite máximo normado de 0.21 ppm o lo que es equivalente a 395 µg/m3, en una hora una vez al año, como protección a la salud de la población susceptible.

4. **Monóxido de Carbono** (CO). El monóxido de carbono similarmente al dióxido de nitrógeno es el residuo de procesos de combustión, sin embargo el monóxido de carbono deriva directamente de de la combustión incompleta de compuestos de carbono, en el Valle de México su principal fuente es la combustión de los vehículos automotores así como la industria.

   El riesgo de la exposición al monóxido de carbono varía desde el efecto de pequeñas cantidades atmosféricas en individuos que padecen deficiencias circulatorias (siendo particularmente susceptibles los enfermos con angina de pecho, así como aquellos con arterioesclerosis), hasta una intoxicación aguda por inhalación de grandes cantidades del contaminante en espacios cerrados y/o en un período corto.

   La concentración de monóxido de carbono, como contaminante atmosférico, no debe rebasar el valor permisible de 11.00 ppm o lo que es equivalente a 12,595 µg/m3 en promedio móvil de ocho horas una vez al año, como protección a la salud de la población susceptible.

5. **Particulas PM10**. Las partículas se definen como cualquier material que existe en estado sólido o líquido en la atmósfera o en una corriente de gas, excepto agua o hielo. Las partículas incluyen polvo, ceniza, hollín, humo y pequeñas partículas de contaminantes. De acuerdo con su diámetro aerodinámico, se clasifican en menores o iguales a 10 µm (PM10). Contienen principalmente materiales de la corteza terrestre y se originan en su mayoría por procesos de desintegración de partículas más grandes. También pueden contener material biológico como polen, esporas, virus o bacterias o provenir de la combustión incompleta de combustibles fósiles.

   Afectan en particular a los sistemas respiratorio y cardiovascular. Toda la población puede ser afectada. Los eventos más documentados son la mortalidad y la hospitalización de pacientes con enfermedad pulmonar obstructiva crónica (EPOC), exacerbación de los síntomas y aumento de la necesidad de terapia en

asmáticos, mortalidad y hospitalización de pacientes con enfermedades cardiovasculares y diabetes mellitus, aumento del riesgo de infarto al miocardio, inflamación de las vías respiratorias, inflamación sistémica, disfunción endotelial y vascular, desarrollo de aterosclerosis, aumento en la incidencia de infecciones y cáncer de pulmón.

Toda la información mencionada anteriormente fue citada directamente de la pagina de la SEMARNAT y Instituto para la Salud Geoambiental: `http://dgeiawf.semarnat.gob.mx:8080/ibi_apps/WFServlet?IBIF_ex=D3_R_AIRE01_01&IBIC_user=dgeia_mce&IBIC_pass=dgeia_mce`, `https://www.saludgeoambiental.org/`.

Con esta explicación recabada, lo que se busca entonces es entonces realizar un análisis de la información existente respecto a las cantidades de estos componentes nocivos para la salud, y en general encontrar si han existido situaciones criticas en el pasado y cuales han sido las acumulaciones mas largas de estos componentes obtenidas como las subsucesiones crecientes mas largas. Con esto en mente, se procede a la elección de la base de datos.

0.6. Elección de la base de datos

Para este análisis se decidió trabajar con la base de datos recabada por el gobierno de la ciudad de México, la base es

”Indice de Calidad del Aire (horarios)” y se puede consultar en el siguiente enlace: `http://www.aire.cdmx.gob.mx/default.php?opc=%27aKBhnmI=%27&opcion=aw==`.

con la cual se trabajaron con los datos en formato .csv (comma separated values).

La base contiene los índices de los componentes nocivos recabados por estaciones atmosféricas distribuidas alrededor del Valle de México, estos índices son calculados de manera diaria promediadas por hora las 24 horas del día, desde 1990 hasta el año 2018.

La base es apropiada puesto que contiene información bien estructurada para realizar el análisis que se requiere, además de que los datos son obtenidos directamente

de un conjunto de estaciones, la recompilación de los datos se realiza físicamente de las mediciones obtenidas para conjuntos de aire que posteriormente son promediadas entre un grupo de estaciones, por lo que difícilmente se presta a obtener variaciones anormales o ocasionadas por alguna clase de ruido.

Para este análisis se trabajara con los datos correspondientes a los años posteriores (e incluyendo) a los del año 2000, posteriormente estos serán manejados de la forma que se requiera para el análisis.

0.6.1. Preprocesamiento de los datos

Los datos son cargados al programa desde los csv en $data\_frames$ sin embargo como se menciono anteriormente estos están dados por hora y por región, puesto que en un comienzo para los años seleccionados se tiene un $data\_frame$ de dimensión 174558 rows × 27 columns.

La forma de reducir los datos fue primero reduciendo las 5 regiones por cada contaminante a una columna por contaminante, esto mediante los promedios de las 5 regiones para cada tipo de contaminante, teniendo ahora un $data\_frame$ de dimensión 174558 rows × 6 columns.

La complicación surge cuando se busca reducir los datos promediados por hora a promediados por día y posteriormente promediados por mes. La forma de realizar esta reducción fue creando nuevos $data\_frames$ para cada reducción, donde en general lo que se realizo fue obtener los promedios de todos los contaminantes para cada fecha disponible y así obtener vectores de acumulación de contaminantes evaluados diariamente.

Análogamente para la reducción a promedios por mes, se extrajeron los meses y años de cada fecha y se calculo el promedio de los contaminantes para cada mes, por cada año, lo cual se guardo en un tercer $data\_frame$. Los $data\_frames$ obtenidos son ahora de dimensiones 6909 rows × 6 columns y 228 rows × 7 columns para promedios diarios y mensuales respectivamente.

0.7. Modelado matemático

Dado que se tienen los datos en $data\_frames$ se piensa en un comienzo que se operara con matrices de datos, sin embargo dado que se realiza el análisis de manera individual para cada contaminante, resulta mas conveniente trabajar con los vectores formados por las columnas de los promedios de cada contaminante.

Así los supuestos serán entonces que cada vector con el que se trabajara representa los datos históricos en el intervalo de tiempo determinado por el $data\_frames$, para algún contaminante seleccionados.

A pesar de que los datos son valores en $\mathbf{R}$, los datos se encuentran dados por vectores distribuidos a través del tiempo en un soporte discreto (días, meses), por esto esta estructura de modelación de los datos históricos es suficiente para la implementación de algoritmos combinatorios.

0.8. Problema algorítmico a resolver

El problema consiste en encontrar una representación que pueda explicar de manera generalizada el comportamiento histórico de los datos de los contaminantes dentro del valle. Una representación que generaliza bien a los datos es encontrar los intervalos mas grandes donde constantemente se empeora la calidad del aire (es decir, la concentración de los contaminantes incrementa de manera constante), y ver cual es el peor de los casos que se ha tenido históricamente.

Debido a que se desea encontrar el intervalo de datos creciente mas grande (normalmente referido como Longest Increasing Subsequence o LIS), el problema resulta entonces en para cada vector de datos de los contaminantes, encontrar la subsecuencia ordenada de menor a mayor donde la subsecuencia es la mas larga posible.

0.8.1. Propuesta de Solución

El problema de encontrar la subsecuencia creciente mas grande es un problema conocido donde la solución tiene una complejidad en tiempo $O(n^2)$ y $O(n)$ en espacio (esto no se supondrá como cierto pues no es este el algoritmo que se empleara y analizara) hecho de la forma intuitiva que consiste en buscar por todo el espacio de soluciones a la subsecuencia creciente.

Existe una solución también mediante arboles binarios de búsqueda que reduce la complejidad de tiempo a $O(nlogn)$, sin embargo esta implementación resulta compleja de implementar, puesto que se hace la búsqueda en un árbol binario lo cual introduce una estructura de datos que requiere su propia implementación dentro del código.

Con lo mencionado anteriormente, se busca entonces desarrollar un algoritmo con complejidad en tiempo al menos tan buena como la del árbol de búsqueda binaria $O(nlogn)$ pero con el empleamiento de estructuras de datos simples como lo son los vectores.

0.8.1.1. Desarrollo del algoritmo

El algoritmo se pensó como una implementación de programación Dinámica con estructuras de vectores, donde en un principio solo se tuviese que recorrer el vector a procesar una sola vez.

Dada la naturaleza del vector que se espera obtener, el algoritmo debe de tener a consideración que las longitudes pueden ir incrementando de distintas maneras en base a que valor del vector se este procesando, por esto la programación dinámica entra en que se tendrá una lista de vectores donde para la iteración actual. Entonces el algoritmo guardara una lista con las subsucesiones que se vayan obteniendo conforme4 se itere sobre el vector.

Dado que se buscan entonces subsucesiones crecientes, se puede realizar un recorte del espacio de estados, donde entonces una ves que un vector sea no factible

se elimine de la lista de subsucesiones formadas, esto reduce considerablemente la ocupación de memoria, lo que reduce la complejidad en memoria del algoritmo.

La forma de definir que una subsucesion deja de ser factible, es cuando esta es alcanzada en longitud por otra de menor distancia en una iteración anterior, puesto que cuando se agregue un valor a la que se eliminara, también se le agrega al otro, por lo que la vieja subsucesion es al menos tan grande como la que la acaba de alcanzar.

Con todo lo anterior mencionado, entonces el algoritmo lo que realiza es iterar sobre el vector de valores del cual se quiere obtener la lis, donde en cada iteración lleve una lista de candidatos de subsucesiones crecientes, la cual se actualizara en cada iteración donde para cada valor que se pueda agregar se agrega donde pueda agregarse, si alguna subsecuencia deje de ser factible y si no se puede agregar ninguna a la lista, se inicializa una nueva subsucesion dentro de la lista con el valor que se esta procesando.

Formalmente el algoritmo realiza los siguientes pasos

1. Inicializar lista de subsucesiones y últimos valores de cada una.

2. Para cada valor en la lista a procesar, se realizan el siguiente paso:

3. Comparar el valor actual $A[i]$ con los valores actuales de la lista de los últimos valores. En base a la comparación, realizar la acción correspondiente:

    - $A[i] < l$ para todos los l en la lista de últimos valores. En este caso, se agrega una nueva subsucesion $[A[i]]$ a la lista de subsucesiones, y se agrega el valor $A[i]$ a la lista de últimos valores.
    - $A[i] > l$ para todos los l en la lista de últimos valores. Se duplica la ultima subsucesion mas grande de las que están en la lista y se le agrega el valor $A[i]$, además de que este también se agrega a la lista de últimos valores.
    - $A[i] > l, A[i] < k$ para l y k en la listas de los últimos valores. Se duplican las subsucesiones donde se puede agregar $A[i]$ (los correspondientes a l en

la lista de últimos valores) al igual que se agrega a la lista de últimos valores, y se eliminan todos las subsucesiones de longitud m correspondiente a la longitud de la subsucesion con el valor agregado $A[i]$ al igual que sus correspondientes últimos valores de la segunda lista.

4. El algoritmo regresa la subsucesion mas larga de la lista de subsucesiones candidatas.

La agregación de la lista de los últimos valores es para simplificar la relación hacia el ultimo valor de cada subsucesion.

0.9. Análisis de correctitud y asintótico

Para el análisis de correctitud del algoritmo primero se revisa que el algoritmo analice todo el espacio de estados. No es difícil de ver que este se lo analiza, pues en un comienzo realiza todas las subsucesiones crecientes posibles del vector que se este procesando. Teniendo todas las subsucesiones el algoritmo regresa la de mayor longitud.

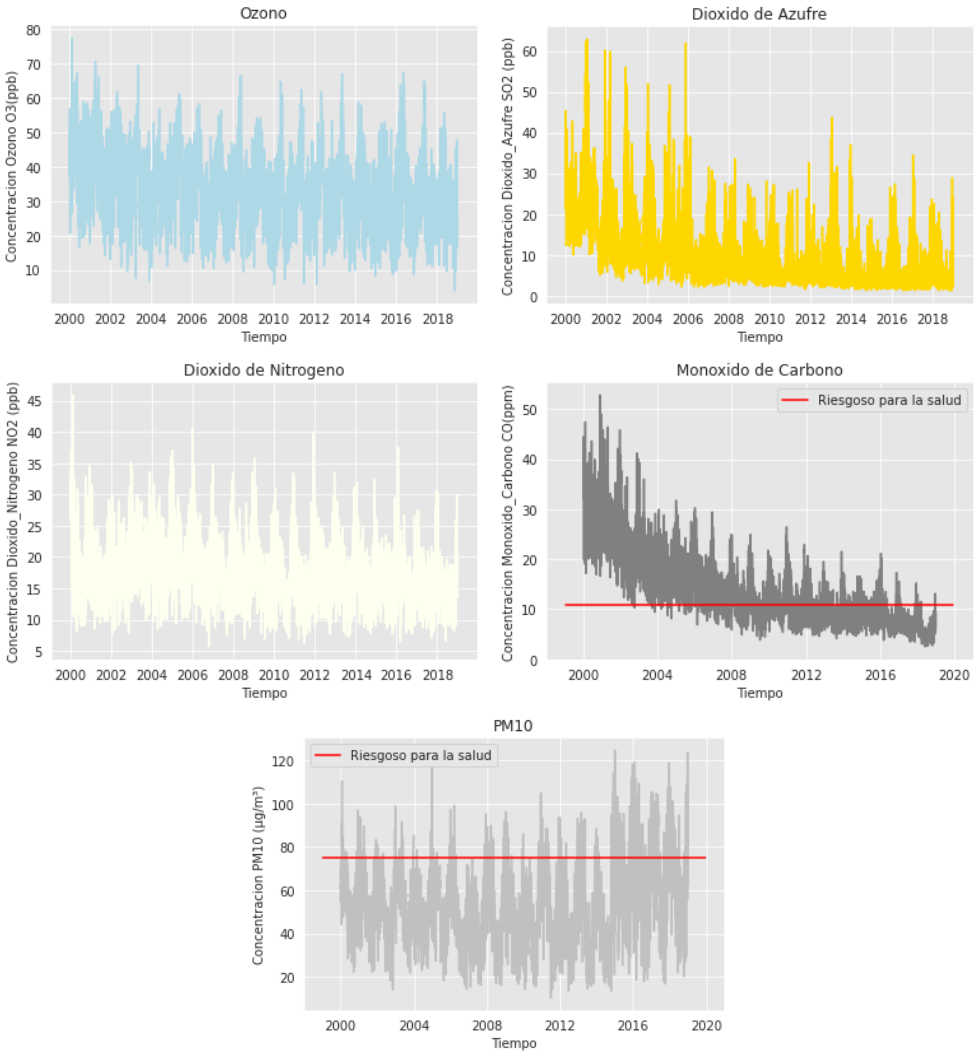
Si bien la correctitud es bastante simple, es una forma suficiente y bastante intuitiva de verificar que en efecto el algoritmo regresa la subsucesion creciente mas grande del vector.

Para el análisis asintótico del algoritmo se parte de que el algoritmo realiza un procedimiento para cada valor del vector, por lo que la complejidad en tiempo es al menos $O(n)$ donde n es el numero de valores en el vector que se este procesando. Dentro de cada iteración, se realizaran comparaciones con las subsucesiones crecientes candidatas hasta esa iteración, debido a la cantidad de subsucesiones es a lo mas $logn$ correspondiente a la longitud del vector entre n listas de longitud 1,(ademas considerar las que no son factibles se eliminan), se realizan en el peor de los casos $logn$ comparaciones, posterior a cada comparación se realizan 2 operaciones por los que en la complejidad son constantes.

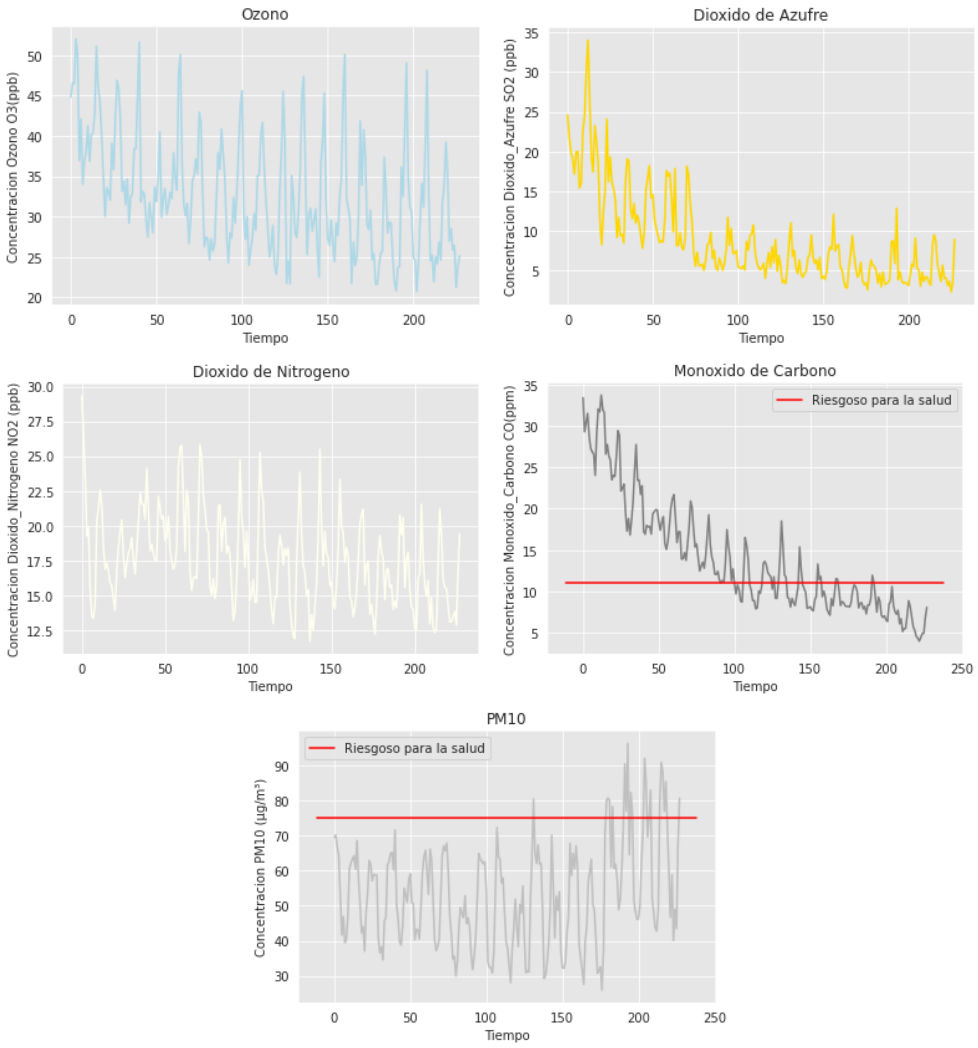
Finalmente para el procesamiento de todo el vector, se realizan entonces $nlogn$ operaciones en el peor de los casos, por esto la complejidad en tiempo del algoritmo es $O(nlogn)$, donde $O(nlogn) < O(n^2)$ la cual corresponde a la implementación intuitiva del algoritmo.

0.9.1. Aplicación a los datos

Primero hay que las reducciones de los datos realizadas para operar con estos. Entonces los datos vistos como vectores para cada contaminante son:



los cuales se ven de manera mas clara cuando se promedian por mes:



Finalmente, con los datos re-escalados, se obtienen las subsucesiones crecientes mas grandes de estos datos:

- **Ozono** (O_3).$(min = 24.63, max = 33.959307795698926, meses\_crecientes : 8)$ Estos rangos indican que si bien se mantuvo una longitud de una subsucesion creciente de al menos 8 meses, se mantuvo por debajo de la mediana del intervalo de los datos, lo cual representa que al menos en el peor de los casos se mantuvo la mayoria del tiempo por debajo de la peor mitad.

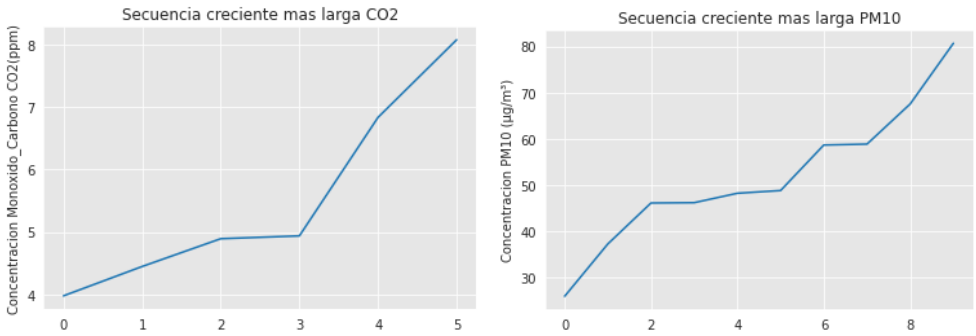
- **Dióxido de Azufre** (SO_2).$(min = 2.85, max = 8.9, meses\_crecientes : 3)$ Para este componente se tuvo un excelente comportamiento pues la subsucesion crecien-

te mas larga fue de 3 meses, en un intervalo bastante bajo comparado con el intervalo general de todos los datos registrados.

- **Dióxido de Nitrógeno** (NO_2).$(min = 11.7, max = 21.21, meses\_crecientes : 14)$ En lo que respecta al Nitrógeno se tuvo una mal comportamiento pues se la subsucesion creciente mas larga se mantuvo dentro de un intervalo bastante grande por 14 meses alcanzando la segunda parte mas grande, afortunadamente este componente en general nunca alcanza niveles críticos para la salud.

- **Monóxido de Carbono** (CO).$(min = 3.98, max = 8.07, meses\_crecientes : 6)$ Al igual que los casos anteriores, en respecto a la acumulación mas grande, esta además se mantuvo totalmente por debajo de la sección critica, sin embargo estuvo cerca de alcanzarla, por lo que respecto a la acumulación mas larga se mantuvo un comportamiento saludable.

- **Partículas PM10**.$(min = 25.98, max = 80.64, meses\_crecientes : 10)$ El PM10 tuvo un comportamiento extremadamente peligroso pues la subsucesion creciente mas larga duro 10 meses en un intervalo bastante amplio abarcando casi todo el intervalo general de los datos históricos y rebaso la marca de la sección nociva para la salud. Este resultado es alarmante para la población puesto que además las partículas PM10 resultan ser de los contaminantes mas letales para la mayoría de las personas.



0.10. Conclusiones y posible trabajo a futuro

Este análisis permitió obtener resultados que no resultan tan intuitivos respecto a los niveles de contaminantes en el aire. Obtener las acumulaciones mas grandes (subsucesiones crecientes mas grandes) permiten obtener una idea de lo que son los tamaños de periodos de tiempos donde en general, se tiene la tendencia de que cada día siguiente sera peor que el anterior, obtener los intervalos y la longitud de estas concentraciones dan una explicación de cuales son los peores casos constantes y que es lo peor que se podría esperar.

Para este análisis se vio entonces que entre los principales contaminantes nocivos para los humanos las peores situaciones son relativamente aceptables para la mayoría de los contaminantes, sin embargo para los 2 mas letales que son el Dióxido de Nitrógeno y las partículas PM10 se han tenido casos extremadamente peligrosos, por lo que se espera que se hagan los cambios necesarios en los estilos de vida de las personas para mejorar la prevención de la emisión de estos contaminantes, así como la prevención de su consumo mediante el aire.

Idealmente se espera que las personas que hayan realizado la lectura de este proyecto se vuelvan mas conscientes en respecto a que actividades realizan que pueden producir estos contaminantes y como evitarlos, puesto que se vio que son extremadamente peligrosos para la salud e históricamente ya se han tenido situaciones criticas, por la salud propia y de los demás se motiva a hacer conciencia sobre el tema tratado.

Para profundizar en el análisis, se pueden realizar primeramente el análisis con los datos a la fecha presente que se considere (en este caso los datos aun no están disponibles para los años 2019 y 2020), pero mas allá de eso un primer paso es considerar que meses son los correspondientes a los mejores y los peores respecto a como es que es que los indicies disminuyen o incrementan respectivamente.

Otra forma de profundizar con algoritmos combinatorios, es obtener para estos datos cuales son las permutaciones de concentraciones que representarían un día ”bueno” conforme a los mejores días obtenidos históricamente, sin embargo este análi-

sis puede requerir bastante poder computacional pues obtener permutaciones de vectores de longitudes tan grandes (al menos 24 horas por día) toma demasiado tiempo y espacio, lo ideal seria encontrar una forma óptima de realizar las permutaciones posiblemente mediante backtrack y recorte del espacio de estados.

Finalmente, otras formas de profundizar fuera del área de las matemáticas discretas son hacer predicciones de los indicies mediante modelos de series de tiempo ARIMA o predicciones de los índices de un contaminante explicado por el resto de los contaminantes mediante modelos de regresión múltiple o incluso redes neuronales.

0.11. Bibliografía

Todas las definiciones e información referenciada se encuentran en los siguientes enlaces:

- http://dgeiawf.semarnat.gob.mx:8080/ibi_apps/WFServlet?IBIF_ex=D3_R_AIRE01_01&IBIC_user=dgeia_mce&IBIC_pass=dgeia_mce
- https://www.saludgeoambiental.org/
- http://www.aire.cdmx.gob.mx/default.php?opc=%27aKBhnmI=%27&opcion=aw==
- https://es.wikipedia.org/wiki/Problema_de_la_subsecuencia_m%C3%A1s_larga

EL notebook donde se realizaron todos los programas y ejecuciones se puede ver en el siguiente enlace:

https://colab.research.google.com/drive/1GW-Au6SROkpA7wMoDKzxAyrDjfe1lQ4_?usp=sharing