Análisis de la Calidad del Aire en el Valle de México

Equipo 3:

- Maria Beatriz Sanchez Díaz
- Leonardo Zurita Martínez
- Jesús Ramseths Echeverría Rivera
- Daniel Alberto Ramos López

Análisis de Datos
BEDU - Santander

Índice

Introducción

La Organización Mundial de la Salud estima que en 2012 aconteció un aproximado de 3.7 millones de muertes prematuras derivadas de enfermedades atribuidas a la contaminación del aire. Esto se refleja en su discurso del 2014: “La contaminación del aire es el riesgo para la salud ambiental número uno del mundo”.

La Ciudad de México, centro político, histórico y financiero de nuestro país cuenta con poco más de 9 millones de habitantes convirtiéndose en la ciudad más poblada de norteamérica, con una densidad de 600,000 habitantes por kilómetro cuadrado. Esto sin contar la totalidad de población que habita la Zona Metropolitana del Valle de México, que se contabiliza por encima de los 21 millones de habitantes.

Los problemas que aquejan a la incomparable urbanización no son pocos, desde crimen organizado, alcoholismo, inundaciones periódicas, fríos intensos y el que concierne a esta investigación, la contaminación del aire. De las majestuosas vistas que los primeros españoles hallaron en aquel valle, no queda nada. Incluso el propio cielo se esconde entre una densa nube de vapores tóxicos.

Siendo conscientes de este problema, desde el año de 1986, se monitorea de forma periódica la calidad del aire en la Zona Metropolitana del Valle de México por medio del Sistema de Monitoreo Atmosférico (SIMAT) y la RAMA (Red Automática de Monitoreo Atmosférico), realizando mediciones de la concentración de los contaminantes criterio cada hora.

Ahora bien, los contaminantes criterio son aquellos que afectan el bienestar y la salud humana, por lo que cuentan con criterios para establecer o revisar límites máximos permisibles por medio de las Normas Oficiales Mexicanas. La concentración de estos contaminantes sirve para indicar la calidad del aire.

Dentro del grupo de contaminantes criterio se encuentran el dióxido de azufre (SO2), el dióxido de nitrógeno (NO2), el monóxido de carbono (CO), el ozono (O3) y las partículas en suspensión (PM10 y PM2.5). Durante nuestra investigación preliminar encontramos información sobre estos contaminantes, como sus afectaciones a la salud y presencia en el Valle de México, que consideramos útil para nuestro análisis:

- *Dióxido de azufre (SO2)*
  Algunas de las fuentes principales de este contaminante son la refinería y termoeléctrica de Tula, el corredor industrial Tula-Vito-Apasco, emisiones volcánicas, quema de combustibles fósiles y manufactura química. Las épocas donde se registra la mayor concentración de este contaminante es entre los meses de noviembre a mayo (temporada seca) y la menor en los meses de junio a octubre (temporada de lluvia).

  Entre sus efectos a la salud se encuentra la irritación de las vías respiratorias, la broncoconstricción y agravantes de enfermedades respiratorias y cardiovasculares.

- *Óxidos de nitrógeno (NOx)*
  Abarca a los compuestos de monóxido de nitrógeno y dióxido de nitrógeno. La mayor fuente de NO2 proviene de la oxidación del NO, además de la quema de combustibles fósiles, biocombustibles y biomasa. Se halla una mayor concentración de estos contaminantes durante los meses noviembre a febrero (fríos y de poca humedad), mientras que de junio a octubre se hallan los valores mínimos (época de lluvia).

  Entre sus efectos a la salud encontramos daños pulmonares, aumento de infecciones respiratorias, e incluso bronquitis y pulmonía.

- *Monóxido de carbono (CO)*
  Su emisión y oxidación en el medio ambiente contribuyen a la formación del CO2. Sus principales emisores provienen de la combustión incompleta de gas natural, propano, gasolina, petróleo, queroseno, madera o cartón. No hay mucha variación en la concentración de este contaminante a lo largo del año.

  Sus efectos adversos a la salud indican que en altas concentraciones inhabilita el transporte de oxígeno hacia las células y que exposiciones prolongadas causan mareos, dolor de cabeza, náuseas, inconsciencia e incluso muerte.

- *Ozono (O3)*
  El ozono presente en la estratosfera es bueno, porque nos protege de los rayos solares, sin embargo su concentración en la troposfera es mala por su impacto en la salud y el ambiente, reduciendo el rendimiento de las cosechas. Su concentración aumenta durante la llamada *temporada de ozono* (de febrero a junio). Además su concentración varía a lo largo del día por la luz solar.

  Tiene efectos adversos en la salud como irritación de las vías respiratorias. En altas concentraciones reduce la función pulmonar, además de que es agravante del asma y de enfermedades pulmonares crónicas.

- *Partículas suspendidas (PM10 y PM2.5)*
  Para los meses invernales de enero y diciembre se hallan las concentraciones más altas y las concentraciones menores se encuentran durante la temporada de lluvias.

  Entre sus efectos a la salud se encuentra que agravan el asma, enfermedades respiratorias y cardiovasculares.

Como se puede observar todos estos contaminantes tienen graves afectaciones a la salud por lo que es necesario no solo monitorear su concentración, sino también obtener con ellos un índice que nos permita clasificar la calidad del aire para así comunicarlo a la población en general. Con este objetivo en mente, en 1982 se diseñó el Índice Metropolitano de la Calidad del Aire (IMECA), cuya metodología transforma a una escala adimensional las concentraciones de los contaminantes criterio; así podemos entender que tan buena o mala es la calidad del aire observando el valor IMECA de dicho contaminante e incluso

compararlo con otro, ya que tienen la misma escala. En la figura 1 se muestra esta escala de clasificación de los valores IMECA.

Intervalo IMECA	Calificativo de la calidad del aire
0 - 50	Buena
51 - 100	Regular
101 - 150	Mala
151 - 200	Muy Mala
> 200	Extremadamente Mala

Figura 1. Clasificación IMECA

Durante nuestra investigación nos dimos cuenta que la información sobre la calidad del aire en la Zona Metropolitana del Valle de México no es tan accesible: existen los datos y están abiertos al público, sin embargo, hay muy poca difusión sobre sus resultados, el último informe anual de la calidad del aire en ciudad de méxico se realizó en el 2018. Por lo tanto el objetivo de este proyecto es primero que nada analizar los datos disponibles para entender la problemática de la contaminación del aire en el Valle de México, y hacer que esta información sea más accesible, para luego proponer posibles soluciones.

Preguntas de investigación

Con la información recabada se plantearon las siguientes preguntas que podrían ayudar a la resolución del problema planteado.

- ¿Cuál es el promedio histórico (considerando los años 2005-2020) de concentración de cada contaminante por zona?
- ¿En qué meses se tiene la máxima concentración de cada contaminante? ¿Existe estacionalidad?
- ¿Hay algún mes de algún año en el que se haya obtenido la calificación de calidad “Extremadamente mala”?
- ¿Cuál es la zona que presenta mayor contaminación, en general?
- ¿Cuál es el contaminante con más concentración por año (del 2005- 2020)?
- ¿Cuál fue y cuándo se registró la concentración más baja durante 2020?
- ¿Cuál fue y cuándo se registró la mayor concentración durante el periodo de confinamiento?
- ¿El periodo de confinamiento coincide con una mejor calidad de aire?
- ¿Cuál fue el contaminante con menores cambios en el período 2005-2020?

Descripción de los datos

Al principio no fue tan sencillo encontrar los datos para investigar la temática de la calidad del aire, pero finalmente decidimos trabajar con las bases de datos del IMECA (Índice Metropolitano de la Calidad del Aire) puesto que nos pareció más informativo y claro usar su escala adimensional; además resume la información de las 34 estaciones de monitoreo existentes en cinco zonas representativas del Valle de México: *Noroeste, Noreste, Centro, Suroeste, Sureste.*

La mayoría de los datos los descargamos en este enlace de la Dirección de Monitoreo Atmosférico, sin embargo los datos del 2019 se encontraban incompletos. Tras una búsqueda más profunda encontramos otro enlace y seleccionando la opción "índice de calidad del aire" accedimos a los datos completos del 2019.

La base de datos estaba conformada por un archivo con extensión .xls para cada año. Cada uno de estos archivos incluye información sobre la fecha y hora de medición, así como las mediciones para cada zona y contaminante criterio. Elegimos trabajar con los datos de los años 2005-2020 porque consideramos que es un buen rango para analizar y observar cambios en el tiempo, así como hacer comparaciones y responder nuestras preguntas de investigación. Una vez que descargamos los datos, los convertimos a archivos CSV para manejarlos más fácilmente en cualquier entorno y con cualquier lenguaje de programación.

Por último es importante resaltar que el diccionario de la base de datos nos indica que los datos nulos se identifican con la etiqueta -99.

Análisis exploratorio de los datos

El conjunto de datos por años desde 2005 hasta el 2018, está organizado de la siguiente manera:

- Fecha
- Hora
- Noroeste Ozono
- Noroeste dióxido de azufre
- Noroeste dioxido de nitrogeno
- Noroeste monóxido de carbono
- Noroeste PM10
- Noreste Ozono
- Noreste dióxido de azufre
- Noreste dioxido de nitrogeno
- Noreste monóxido de carbono
- Noreste PM10
- Centro Ozono
- Centro dióxido de azufre
- Centro dioxido de nitrogeno
- Centro monóxido de carbono
- Centro PM10
- Suroeste Ozono
- Suroeste dióxido de azufre
- Suroeste dioxido de nitrogeno
- Suroeste monóxido de carbono
- Suroeste PM10
- Sureste Ozono
- Sureste dióxido de azufre
- Sureste dioxido de nitrogeno
- Sureste monóxido de carbono
- Sureste PM10

En los años 2019 y 2020 se agregaron mediciones de partículas menores a 2.5 micras (PM2.5), también se modificó el nombre de las columnas:

- Fecha
- Hora
- NOO3
- NOSO2
- NONO2
- NOCO
- NOPM10
- NOPM2
- NEO3
- NESO2
- NENO2
- NECO
- NEPM10
- NEPM2
- CEO3
- CESO2
- CENO2
- CECO
- CEPM10
- CEPM2
- SOO3
- SOSO2
- SONO2
- SOCO
- SOPM10
- SOPM2
- SEO3
- SESO2
- SENO2
- SECO
- SEPM10
- SEPM2

Las mediciones de contaminantes en valor IMECA se registran cada hora los 365 días (o 366 en año bisiesto). Existen datos “-99” que, como se mencionó antes, corresponden a valores faltantes.

El conjunto total de datos cuenta con 140277 filas, cada una equivalente a una medición horaria.

En cuanto a las variables o columnas, se puede resumir que trabajaremos únicamente con 5 contaminantes:

- Ozono (O3)
- Dióxido de azufre (SO2)
- Dióxido de nitrógeno (NO2)
- Monóxido de carbono (CO)
- Partículas menores a 10 micras (PM10)

Y 5 zonas:

- Zona Sureste
- Zona Suroeste
- Zona Centro
- Zona Noreste
- Zona Noroeste

Para el análisis de la limpieza de datos se optó por algunos métodos que redujeron de forma considerable la cantidad de información. Esto se describe a continuación.

Limpieza de datos

Uno de los pasos más importante al momento de trabajar con bases de datos es el paso anterior al procesamiento, que requiere de una limpieza de información. Nos topamos con dificultades que se presentaron por la forma en la que los datos fueron almacenados en el archivo CSV: columnas completas sin información, filas adicionales vacías y cambios de nombres de las columnas entre los años, como se comentó con anterioridad.

Durante la limpieza de datos, al momento de revisar el tamaño de los Data Frame por año, descubrimos que no todos tenían el mismo número de filas, había unos que tenían más. Pero esto no es ningún error, más bien algo de sentido común. Cada cuatro años, tenemos un año bisiesto, es decir, los años que tienen filas “de más” realmente corresponde a 24 exactamente, que serían las mediciones de un día completo, el 29 de Febrero.

Limpieza de valores NaN

Las columnas y 21 filas vacías fueron eliminadas durante este procedimiento. Estudiando el Data Frame, hallamos valores faltantes (-99 equivalentes a NaN). La estrategia que se eligió fue sustituir los valores de tipo NaN por el promedio de la columna correspondiente, de esta manera facilitamos la programación y no fue necesario eliminar filas que podrían dificultar el procesamiento en programación.

Una vez casi concluida la fase de limpieza de datos descubrimos que existían filas de fechas y horas con valores NaN, lo cual no debería ser posible porque los documentos csv estaban completos en todas las fechas. Realizando una análisis más profundo descubrimos que este fenómeno se daba en el cambio de año del 2013 al 2014, donde el archivo csv del 2013 tenía exactamente 21 filas vacías, pero estas filas eran adicionales a las que ya estaban contempladas en el año, por lo cual no se perdió en ningún momento información y solo las descartamos.

Transformación de datos

Revisando el tipo de dato por medio de la función dtypes, observamos que la fecha y hora estaban clasificadas de forma incorrecta, con valores de tipo string y float, respectivamente.

Se realizó un casting a la fecha transformando a tipo datetime especificando el uso de nanosegundos debido que al utilizar en milisegundos nos marcaba error. Consideramos que la hora en realidad no nos será tan útil en análisis posteriores, ya que pensamos agrupar por mes, sin embargo la cambiamos a tipo int para no generar confusiones.

Después de estos procesos de limpieza y transformación, guardamos los dataframes en archivos csv para poder utilizarlos en los siguientes análisis.

Como último paso del proceso de transformación de datos, decidimos calcular promedios mensuales de cada contaminante y zona, reduciendo así la cantidad de datos y quedándonos con valores representativos que nos puedan facilitar el análisis. Para llevar a cabo esta transformación usamos la función resample, agrupando por mes y calculando el promedio. Guardamos estos nuevos dataframes en archivos csv.

Resultados preliminares

A continuación presentamos unos breves resultados obtenidos a partir de lo aprendido durante este módulo. Estos datos son preliminares, es decir carecen de análisis estadístico profundo.

¿Hay algún mes de algún año en el que se haya obtenido la calificación de calidad “Extremadamente mala”?
A partir de una tabla y un filtrado sencillo fue posible descubrir que no existía ningún contaminante que llegará a ese nivel, en el lapso de tiempo 2005-2020. Esto coincide con un análisis descrito en el reporte anual “Calidad del aire en la ZMVM.Reporte anual 2018”, existe una tendencia decreciente en IMECA para los contaminantes de dióxido de azufre, dióxido de nitrógeno y dióxido de carbono. Además de que el resto de contaminantes se han mantenido en niveles tolerables.

¿Cuál es el promedio histórico (considerando los años 2005-2020) de concentración de cada contaminante por zona?

```
noroeste_ozono                  28.90
noroeste_dioxido_de_azufre       9.38
noroeste_dioxido_de_nitrogeno   17.48
noroeste_monoxido_de_carbono    10.09
noroeste_pm10                   55.74
noreste_ozono                   31.22
noreste_dioxido_de_azufre        7.53
noreste_dioxido_de_nitrogeno    16.09
noreste_monoxido_de_carbono      9.96
noreste_pm10                    69.31
centro_ozono                    25.97
centro_dioxido_de_azufre         6.44
centro_dioxido_de_nitrogeno     18.36
centro_monoxido_de_carbono      10.90
centro_pm10                     52.78
suroeste_ozono                  36.57
suroeste_dioxido_de_azufre       5.01
suroeste_dioxido_de_nitrogeno   15.20
suroeste_monoxido_de_carbono     8.51
suroeste_pm10                   40.04
sureste_ozono                   33.09
sureste_dioxido_de_azufre        4.68
sureste_dioxido_de_nitrogeno    15.97
sureste_monoxido_de_carbono     10.08
sureste_pm10                    51.64
dtype: float64
```

Observando los promedios históricos, podemos determinar con claridad que entre la Zona Norte y la Zona Sur existe una diferencia significativa en las concentraciones de los diferentes contaminantes.

¿Cuál es la zona que presenta mayor contaminación, en general?

A partir de los datos obtenidos anteriormente podemos responder igualmente a esta pregunta

- Ozono: Suroeste, Sureste, Noreste
- Dióxido de azufre: Noroeste, Noreste, Centro
- Dióxido de nitrógeno: Centro, Noroeste, Noreste
- Monóxido de carbono: Centro, Noroeste, Sureste
- PM10: Noreste, Noroeste, Centro

A partir de esta información podemos concluir que la Zona Centro es la que presenta en datos históricos los mayores valores IMECA.

¿Cuál es el contaminante con más concentración por año (del 2005- 2020)?

```
fecha
2005-12-31    noreste_pm10
2006-12-31    noreste_pm10
2007-12-31    noreste_pm10
2008-12-31    noreste_pm10
2009-12-31    noreste_pm10
2010-12-31    noreste_pm10
2011-12-31    noreste_pm10
2012-12-31    noreste_pm10
2013-12-31    noreste_pm10
2014-12-31    noreste_pm10
2015-12-31    noreste_pm10
2016-12-31    noreste_pm10
2017-12-31    noreste_pm10
2018-12-31    noreste_pm10
2019-12-31    noreste_pm10
2020-12-31    noreste_pm10
Freq: A-DEC, dtype: object
```

De las concentraciones de contaminantes por año podemos observar que las partículas menores a 10 micras son las que tuvieron un valor IMECA más alto en todos los años. Este dato es de interés porque según informes anteriores estos valores van en descenso, pero de cualquier modo siguen siendo altos.

A futuro

Una vez finalizado todo este proceso de limpieza de datos llega una parte importante a modo de análisis introspectivo de la información limpia. De las preguntas de investigación propuestas desde un principio, las dividiremos de la siguiente manera:

Preguntas contestadas:

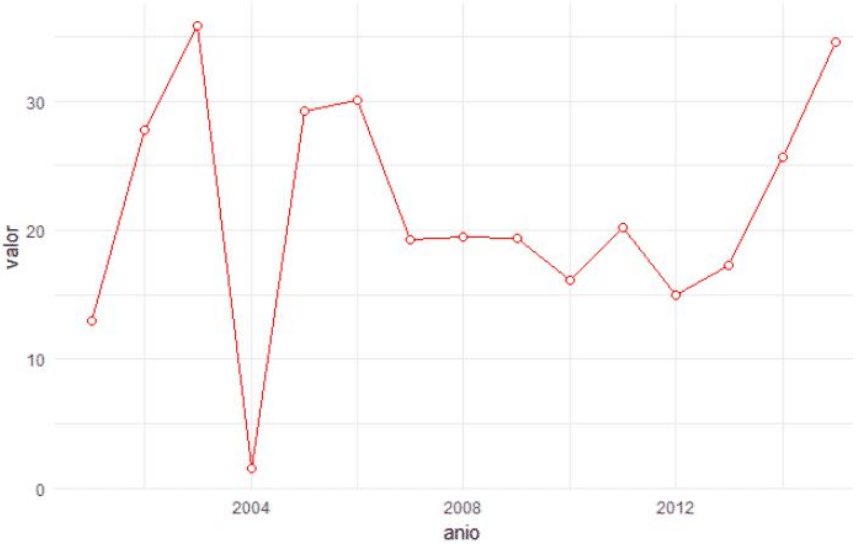
- ¿Cuál es la zona que presenta mayor contaminación, en general?
- ¿Cuál es el contaminante con más concentración por año (del 2005- 2020)?
- ¿Cuál es el promedio histórico (considerando los años 2005-2020) de concentración de cada contaminante por zona?
- ¿Hay algún mes de algún año en el que se haya obtenido la calificación de calidad “Extremadamente mala”?

Preguntas por contestar:

- ¿En qué meses se tiene la máxima concentración de cada contaminante?
- ¿Cuál fue y cuándo se registró la concentración más baja durante 2020?
- ¿Cuál fue y cuándo se registró la mayor concentración durante el periodo de confinamiento?
- ¿El periodo de confinamiento coincide con una mejor calidad de aire?
- ¿Cuál fue el contaminante con menores cambios en el período 2005-2020?

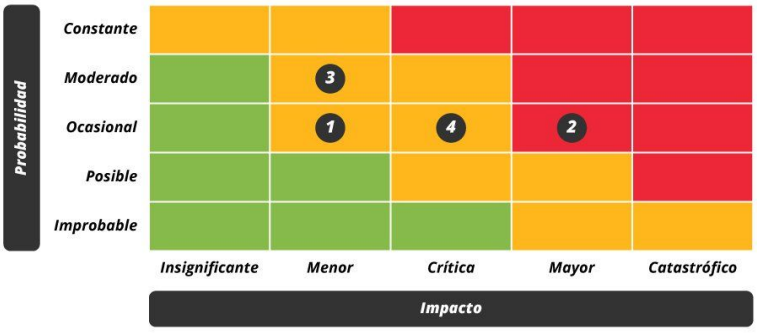
La forma en la que contestamos la información, en resumen, podemos contestar las preguntas, pero las respuestas serían mejor interpretadas por medio de gráficas que nos pueden dar un enfoque más preciso y concreto de la información que estamos obteniendo para identificar y magnificar el impacto de estos datos.

Gráfica de puntos

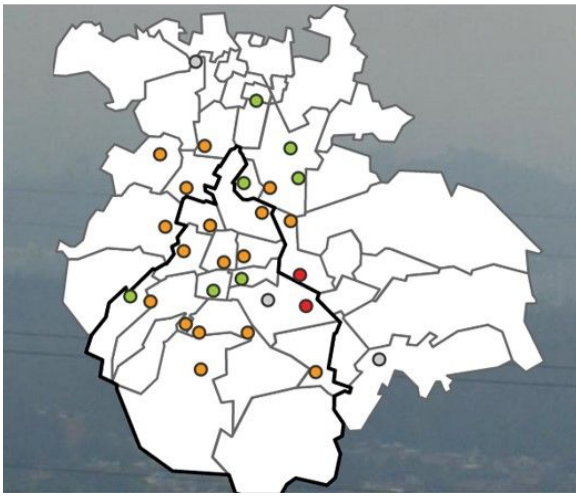


Usaremos una gráfica de líneas para visualizar con facilidad los valores promedios por días,años y meses. De esta manera es posible analizar de forma visual, la evolución de los datos a través del tiempo, como en una Serie de tiempo.

Mapa de calor



Esta gráfica permitirá visualizar de forma sencilla el parámetro de IMECA correspondiente por color de día, durante el 2020. Es útil conocer qué parámetro de calidad de aire se presentó durante los meses de confinamiento.



Mapa geográfico comparativo con puntos.
Estos mapas comparativos nos permitirán enseñar de forma clara cómo es que unas zonas se hallan más o menos contaminadas según el parámetro IMECA.

La propuesta para presentar la información es realizar mapas interactivos utilizando el paquete Folium y datos geográficos de las zonas IMECA. Estas visualizaciones serán muy útiles para hacer comparaciones entre las zonas y tener una mejor percepción de cómo se distribuye la concentración de contaminantes.

Conclusiones

Una vez finalizado este módulo nos percatamos de la importancia del procesamiento de datos, de forma curiosa, cuando vemos una gráfica no pensamos en todo el trabajo que implica obtener esa información. Debido a que la ciencia de datos se perfila como una de las áreas de más crecimiento a futuro, es vital importancia entender que no toda la información es útil. En la era de la tecnología y la información debemos aprender a filtrar datos, tratar con los datos y desarrollar la habilidad de realizar una limpieza en ellos.

En algunas ocasiones, el acceso a la información no es sencillo como nosotros pudimos comprobar, lo cual resalta más la importancia de la información. Un ejemplo claro de esto nos sucedió con unos datos corruptos del año 2019, de no ser porque existía una dependencia que nos permitía hacer consultas de forma específica esa información se hallaría aún extraviada.

Entendemos la importancia que tiene el aire y su calidad,en la vida de la gente de la ZMVM, se habla de los efectos nocivos en la salud que puede tener este, lo cual nos hace reflexionar, en el contexto actual de una pandemia global, ¿la contaminación del aire en nuestras ciudades podría estar afectando de alguna forma las hospitalizaciones?, o ¿será que la pandemia ha sido un factor determinante en la mejora del aire?. No responder de forma clara hasta este punto alguna de estas preguntas, pero es claro que los datos que poseemos podrían ayudar a contestar alguna de estas cuestiones.

Fuentes Bibliográficas

Reporte de calidad del aire

SEDEMA. (2018). INFORME ANUAL CALIDAD DEL AIRE 2018 CIUDAD DE MÉXICO. Ciudad de México: SEDEMA.

Estadísticas

Statista. 2020. COVID-19: change in air pollution Mexico City | Statista. [online] Disponible en: <https://www.statista.com/statistics/1123848/change-no2-air-pollution-mexico-city/> [Revisado 9 Marzo 2021].

OMS

Who.int. 2018. *Household air pollution and health*. [online] Disponible en: <https://www.who.int/en/news-room/fact-sheets/detail/household-air-pollution-and-health> [Revisado 9 Marzo 2021].

Royal Geographic Society

21st Century Challenges. 2015. *Air pollution*. [online] Disponible en: <https://21stcenturychallenges.org/air-pollution/> [Revisado 9 Marzo 2021].

INEGI

Inegi.org.mx. 2021. *México en cifras*. [online] Disponible en: <https://www.inegi.org.mx/app/areasgeograficas/> [Revisado 9 Marzo 2021].

Dirección de monitoreo atmosférico

Aire.cdmx.gob.mx. 2021. *Dirección de Monitoreo Atmosférico*. [online] Disponible en: <http://aire.cdmx.gob.mx/default.php> [Revisado 9 Marzo 2021].