Tecnológico de Monterrey

Etapa 4. Informe Final

Aplicación de Métodos Multivariados en Ciencia de Datos

Alumnos

Fernando Soto - A01252884

Valeria Sada - A00837046

Leonardo De Regil - A00837118

Rogelio Coria López - A01733314

28 de Noviembre del 2024

Profesoras

Blanca Rosa Ruiz Hernandez

Monica Guadalupe Elizondo Amaya

Resumen

El presente informe aborda la problemática de la contaminación del aire en Monterrey, donde actividades humanas e industriales generan impactos significativos en la calidad ambiental y en la salud pública. De acuerdo con datos de la Organización Mundial de la Salud (OMS), la contaminación atmosférica contribuye a millones de muertes prematuras anuales a nivel global, lo que refuerza la urgencia de implementar medidas efectivas. Este estudio se realiza en colaboración con el Sistema Integral de Monitoreo Ambiental (SIMA) y la Dirección de Gestión de Calidad del Aire, entidades que operan 15 estaciones de monitoreo en Monterrey, evaluando contaminantes clave como PM10, PM2.5, O3, SO2, NO2 y CO.

Objetivo

El objetivo del proyecto es analizar cómo las concentraciones de contaminantes varían espacial y temporalmente en las estaciones de monitoreo de Cadereyta, Santa Catarina y Juárez, evaluando la influencia de fuentes industriales y locales. Este análisis busca generar estrategias que mitiguen eventos de alta contaminación y protejan la salud pública.

Metodología

La metodología incluyó:

1. **Preparación de datos**: Imputación de valores faltantes con información histórica y métodos como "último valor observado".
2. **Análisis estadístico**: Evaluación de correlaciones, medias y varianzas para identificar patrones y distribuciones de contaminantes.
3. **Modelos predictivos**: Uso de ARIMA y técnicas de minería de datos para predecir niveles de contaminación y analizar relaciones entre estaciones.

Resultados Principales

- Las estaciones muestran un sesgo a la derecha en la distribución de los contaminantes, indicando la presencia de valores extremos. Cadereyta destaca por sus mayores concentraciones promedio de CO, NO y NOX, atribuibles a la refinería cercana.
- Los análisis de correlación y clusters revelaron patrones temporales y espaciales similares entre las estaciones, aunque con variaciones en magnitud.
- Los modelos predictivos lograron un desempeño aceptable en la estimación de contaminantes como PM10, aunque con limitaciones para SO2 debido a su baja variabilidad.

Conclusiones

El estudio confirma que las emisiones industriales, junto con factores meteorológicos y locales, influyen significativamente en la calidad del aire. La identificación de patrones permite anticipar eventos críticos, proporcionando una base sólida para políticas públicas que promuevan un entorno urbano más saludable. Aunque los modelos presentan limitaciones, ofrecen una herramienta útil para la gestión ambiental en Monterrey.

Este análisis subraya la necesidad de continuar fortaleciendo la infraestructura de monitoreo y la investigación interdisciplinaria para abordar de manera integral la problemática de la contaminación del aire.

Introducción y Justificación

Las actividades humanas, como las realizadas en los hogares, la industria, el transporte, los comercios y los servicios, junto con el manejo de residuos de todo tipo, han generado impactos negativos en el medio ambiente y efectos perjudiciales para la salud humana. Estas actividades producen una variedad de contaminantes atmosféricos que contribuyen al deterioro de la calidad del aire, afectando no solo al entorno, sino también a la salud de las personas, especialmente en áreas urbanas densamente pobladas.

Según datos de la Organización Mundial de la Salud (OMS) y del Programa de las Naciones Unidas (PNUD), nueve de cada diez personas en el mundo respiran aire contaminado, lo que provoca anualmente la muerte prematura de aproximadamente 7 millones de personas a nivel global. Esta alarmante cifra subraya la urgencia de abordar los problemas de contaminación del aire y la necesidad de implementar medidas efectivas para mejorar la calidad del ambiente y la salud de la población.

Para este proyecto estaremos trabajando con Sima y La Dirección de Gestión de Calidad del Aire. El Sistema Integral de Monitoreo Ambiental (SIMA) tiene como objetivo evaluar la calidad del aire monitoreando las concentraciones de los contaminantes atmosféricos a las que se encuentra expuesta la población y, bajo condiciones adversas, advertirle sobre los episodios de altos índices de contaminación atmosférica. El SIMA está integrado por 15 estaciones fijas localizadas en el área metropolitana de Monterrey.

La Dirección de Gestión de Calidad del Aire trabaja en colaboración con SIMA para diseñar estrategias que permitan identificar patrones en los niveles de contaminación y promover políticas de regulación y mitigación. Esta sinergia resulta clave para enfrentar la problemática de la contaminación en un entorno urbano y con condiciones climáticas particulares como las de Monterrey.

Problemática y Objetivo

El problema que se nos presentan SIMA y La Dirección de Gestión de Calidad del Aire es realizar una análisis de datos relacionada con los principales factores que influyen en

la calidad del aire: Conocimiento de la naturaleza de contaminantes que influyen en la calidad del aire y sus interrelaciones con el medio.

Nuestro Objetivo

Nuestro objetivo es evaluar la calidad del aire en Monterrey, enfocándonos en cómo las concentraciones de contaminantes atmosféricos varían entre las estaciones de monitoreo de Cadereyta, Santa Catarina y Juárez. Este objetivo busca determinar las influencias de fuentes industriales y locales en los niveles de contaminación, además de identificar patrones temporales y espaciales que permitan predecir y mitigar eventos de alta contaminación.

Justificación del Objetivo

La calidad del aire en Monterrey representa una preocupación creciente debido a sus efectos adversos en la salud de la población y el medio ambiente. Este estudio se justifica por la necesidad de comprender cómo las emisiones industriales (específicamente la refinería de Cadereyta) afectan de manera diferencial a las distintas áreas de la ciudad, específicamente en Cadereyta, Santa Catarina y Juárez.

Además, la identificación de patrones espaciales y temporales de contaminación resulta crucial para el diseño de estrategias de mitigación y políticas públicas más efectivas. Al predecir eventos de alta contaminación, será posible implementar medidas proactivas que protejan la salud pública y reduzcan los impactos negativos sobre el entorno urbano. Este enfoque no solo contribuye al bienestar de los habitantes de Monterrey, sino que también fortalece las capacidades de monitoreo y gestión ambiental en la región.

Preparación de la Base de Datos

Imputación de Datos

Para completar los registros de datos faltantes del año 2024, se emplearon dos enfoques de imputación de datos basados en información histórica. En primer lugar, se utilizaron los registros de los años 2022 y 2023 para llenar las fechas faltantes del 2024. Esta estrategia permitió aprovechar patrones y tendencias de años previos que pudieran ser consistentes con el comportamiento esperado en 2024.

Para aquellos datos que permanecieron ausentes tras esta imputación, se utilizó una técnica de imputación por último valor observado, en la cual se asigna el valor más reciente registrado antes del dato faltante. Este método de imputación secuencial permite mantener la coherencia temporal en los datos y minimizar las distorsiones que podrían afectar el análisis.

Descripción de las Variables

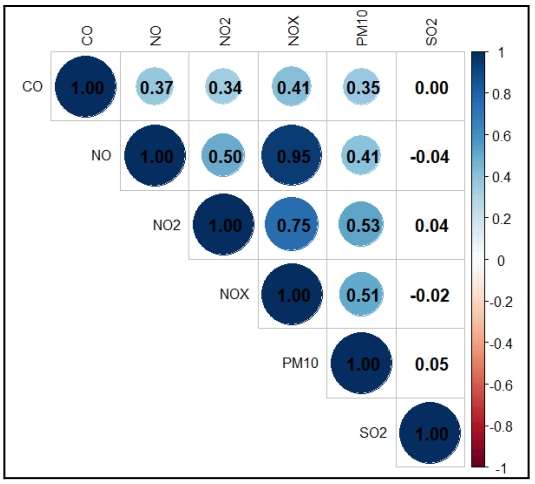
La calidad del aire se mide mediante estaciones de monitoreo que capturan y analizan las concentraciones de contaminantes atmosféricos. Estas estaciones utilizan sensores para capturar cada tipo de contaminante. En la medición de la calidad del aire se toman en cuenta las concentraciones de contaminantes específicos, estos siendo:

PM10	PM2.5	O_3
SO_2	NO_2	CO

También se toman en cuenta factores meteorológicos como presión atmosférica, humedad, temperatura, y velocidad del viento dado que estos pueden influir en la dispersión y acumulación de los contaminantes.

Análisis de Correlación

Aquí se muestra un grafo de correlación de los datos de Cadereyta. En los resultados de correlación, se observa que NO, NO2, y NOX están fuertemente correlacionados entre sí, especialmente NO y NOX (0.95), lo que sugiere que comparten una fuente común o están relacionados en el proceso de contaminación. CO tiene una correlación moderada con NOX (0.41) y PM10 (0.35), indicando una relación entre los contaminantes. Sin embargo, SO2 muestra correlaciones muy débiles o nulas con el resto de las variables, lo que sugiere que no está fuertemente relacionado con los otros contaminantes en este conjunto de datos.



En general, todas las correlaciones de las variables de las estaciones, son muy similares, por lo que mejor optamos hacer un análisis en las medias y varianzas de cada variables y compararlas con los resultados de cada una de las estaciones.

Análisis de Medias y Varianzas

A continuación, se muestran las tablas con los resultados correspondientes a cada una de las estaciones y sus variables con métricas como la media, la varianza, los cuartiles, así como los valores mínimos y máximos. Este análisis es fundamental para poder comparar los resultados de cada una de las estaciones de nuestro análisis.

Análisis en Cadereyta

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Variance	STD	Mode
CO	-0.13	0.84	1.53	1.61	2.08	12.78	1.074158	1.036416	0.74

NO	0.6	3	3.9	8.726	5.8	370.8	316.611458	17.793579	2.6
NO2	0.1	4.8	7.3	10.47	13.3	72.1	72.000631	8.485319	4.5
NOX	0.9	8.2	11.3	19.07	19.2	378	537.281856	23.179341	8.4
PM10	2	42	58	70.81	82.75	999	2769.956266	52.630374	46.0
SO2	0.5	3.3	4.1	6.167	5.1	222.9	78.763600	8.874886	4.2

Los datos sugieren que mientras se observan valores moderados para la mayoría de los contaminantes (basado en sus medias y cuartiles), tenemos algunos valores extremos, especialmente para NO, NOX y PM10.

- **CO**: *Mean* y *median* tienen valores cercanos, sugiriendo una distribución uniforme. El máximo de 12.78 indica un extremo muy alto, pero dado a que el tercer cuartil es 2.08, se considera que este extremo es una irregularidad.
- **NO**: *Mean* y *median* tienen una diferencia significante (8.729 y 3.9 respectivamente), indicando un sesgo a la derecha o valores altos de NO. El valor máximo de 370.8 indica un valor extremadamente alto comparado a el 3er cuartil.
- **NO2**: Al igual que NO, se observa un sesgo a la derecha, a diferencia este no es tan dramático dado a que su máximo es un 72.1, indicando algunos valores extremos.
- **NOX**: Sesgado a la derecha y con valores extremos, se indica que hay valores muy altos presentes de NOX, lo cual lo hace un contaminante abundante más veces que no.
- **PM10**: El *mean* y el *median* indican que hay una distribución sesgada a la derecha, con algunos valores altos de PM10 ocurriendo de vez en cuando. El máximo nos indica que estos valores altos son muy altos.
- **SO2**: Algo de sesgo a la derecha, con valores extremos de SO2 presentes ocasionalmente.

Análisis en Santa Catarina

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Variance	STD	Mode
CO	0.05	0.7	1.51	1.531	2.26	6.52	0.731032	0.855004	0.54
NO	0.8	3.1	4.9	14.24	11.5	362.7	713.324658	26.708138	2.70
NO2	0.2	10.3	15.8	18.45	24.1	92.5	120.769140	10.989501	0.30
NOX	0.6	13.7	21	32.6	36.7	412.7	1184.356458	34.414480	13.40
PM10	3	37	55	68.11	83	802	2576.997780	50.764139	44.00

SO2	0.5	4.7	5.4	5.936	6.6	54.9	7.605562	2.757818	5.30

Interpretación: Los datos sugieren que, aunque los niveles de la mayoría de los contaminantes son moderados (basado en sus medias y cuartiles), existen valores extremos, especialmente para NO, NOX y PM10.

- **CO**: La media y la mediana tienen valores cercanos (1.531 y 1.51), lo cual sugiere una distribución uniforme de los niveles de CO. El máximo de 6.52 indica un valor alto, pero dado que el tercer cuartil es 2.26, este máximo parece ser una irregularidad.
- **NO**: La media (14.24) y la mediana (4.9) muestran una diferencia significativa, indicando un sesgo a la derecha o valores altos de NO. El valor máximo de 362.7 es extremadamente alto en comparación con el tercer cuartil (11.5), sugiriendo la presencia de picos considerables.
- **NO2**: Al igual que NO, muestra un sesgo a la derecha, aunque menos pronunciado, con una media de 18.45 y una mediana de 15.8. Su máximo de 92.5 indica algunos valores extremos de NO2.
- **NOX**: Exhibe un sesgo a la derecha y valores extremos, con una media (32.6) notablemente superior a la mediana (21). El valor máximo de 412.7 indica niveles muy altos de NOX en algunas ocasiones, lo que sugiere una abundancia frecuente de este contaminante.
- **PM10**: La media (68.11) y la mediana (55) reflejan un sesgo a la derecha, con algunos valores altos de PM10 ocasionales. Su máximo de 802 es extremadamente alto, lo que indica la ocurrencia de episodios de alta concentración de partículas.
- **SO2**: Muestra algo de sesgo a la derecha, con una media de 5.936 y una mediana de 5.4. El máximo de 54.9 sugiere la presencia ocasional de valores extremos de SO2.

Análisis en Juárez

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Variance	STD	Mode
CO	0	0.76	1.35	1.375	1.79	8.25	0.560639	0.748758	0.65
NO	0.5	3	5.3	15.49	11.9	400.3	967.638078	31.106882	2.6
NO2	0	5.2	9.4	13.39	18.5	88	131.456576	11.465451	4.7
NOX	1	8.4	15	28.72	31.2	458.4	1530.269031	39.118653	6.0
PM10	2	40	60	73.3	90	729	2744.534375	52.388304	52.0
SO2	0.6	2.6	3.1	4.188	4.1	178.3	20.026629	4.475112	2.7

Interpretación: Los datos indican que, si bien los niveles de la mayoría de los contaminantes son generalmente moderados (según las medias y los cuartiles), se observan valores extremos en NO, NOX y PM10.

- **CO**: La media y la mediana son cercanas (1.375 y 1.35), lo cual sugiere una distribución uniforme de los niveles de CO. El valor máximo de 8.25 es alto, pero como el tercer cuartil es 1.79, este máximo parece ser una irregularidad.
- **NO**: La diferencia significativa entre la media (15.49) y la mediana (5.3) sugiere un sesgo a la derecha o valores altos de NO. El valor máximo de 400.3 es extremadamente alto en comparación con el tercer cuartil (11.9), indicando la presencia de picos pronunciados.
- **NO2**: Similar a NO, presenta un sesgo a la derecha, aunque menos marcado, con una media de 13.39 y una mediana de 9.4. Su máximo de 88 sugiere la presencia de valores extremos.
- **NOX**: Exhibe un sesgo a la derecha y valores extremos, con una media (28.72) superior a la mediana (15). El valor máximo de 458.4 indica niveles muy altos de NOX en algunas ocasiones, lo que sugiere una abundancia frecuente de este contaminante.
- **PM10**: La media (73.3) y la mediana (60) reflejan un sesgo a la derecha, con la ocurrencia ocasional de valores altos de PM10. Su máximo de 729 es extremadamente alto, lo que indica la presencia de episodios con alta concentración de partículas.
- **SO2**: Muestra algo de sesgo a la derecha, con una media de 4.188 y una mediana de 3.1. El valor máximo de 178.3 sugiere la presencia ocasional de valores extremos de SO2.

En general, los datos de las tres estaciones muestran patrones similares en los contaminantes, con niveles medios y cuartiles relativamente moderados, pero con la presencia de valores extremos en varios contaminantes, especialmente en **NO**, **NOX** y **PM10**, que sugieren episodios de alta concentración de estos contaminantes en ciertas ocasiones.

A continuación, se presentará una serie de boxplots que permiten comparar las concentraciones de diferentes contaminantes entre tres estaciones de monitoreo: Cadereyta, Santa Catarina y Juárez. Cada gráfico representa una variable ambiental específica (como CO, NO, NO2, NOX, PM10 y SO2) y muestra cómo varían las concentraciones de estos contaminantes en las tres estaciones de monitoreo. La comparación visual de los boxplots facilita la identificación de posibles patrones, tendencias y diferencias significativas en los niveles de contaminación entre las ubicaciones. Los gráficos de los boxplots se encuentran adjuntos en las referencias (link de Drive) de este documento.

Conclusión e Interpretación

El análisis muestra un sesgo derecho en las distribuciones de las variables, destacando valores extremos que reflejan periodos de alta exposición a contaminantes en Monterrey. Cadereyta presenta mayores concentraciones promedio de contaminantes como CO, NO y NOX debido a su cercanía a una refinería, confirmando el impacto de fuentes industriales. Santa Catarina y Juárez también exhiben niveles significativos de PM10 y otros contaminantes, posiblemente asociados a emisiones locales o condiciones específicas. Los boxplots revelaron patrones similares entre las estaciones, aunque las diferencias en magnitud y frecuencia de valores extremos serán clave para análisis futuros.

Modelación y Validación

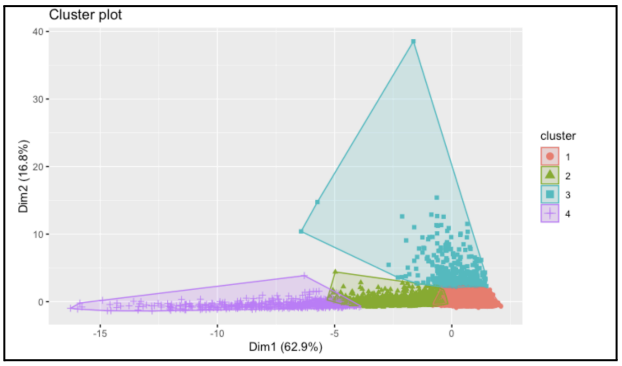
Relaciones de Interdependencia

Se intentó realizar un análisis factorial sobre los datos por estación, por lo que se realizaron pruebas de adecuación muestral de Kaiser-Meyer-Olkin (KMO) para los contaminantes registrados en cada estación. Los resultados arrojaron valores insuficientes que indicaron que los datos no eran aptos para realizar un análisis factorial.

A pesar de esta limitación, se procedió a ejecutar el análisis factorial con el objetivo de explorar posibles patrones o relaciones latentes entre los contaminantes. Los resultados del análisis factorial confirman las advertencias de las pruebas KMO: los modelos generados no lograron explicar adecuadamente la variabilidad de los datos, y los factores extraídos no presentaron una interpretación clara o consistente. Dado el desempeño limitado de esta metodología, se decidió explorar alternativas más adecuadas para analizar y modelar los datos de contaminación, adaptándonos mejor a las características de las variables y al objetivo del estudio. El análisis de interdependencias entre variables se puede abordar mediante métodos estadísticos y técnicas de minería de datos. En este caso, los clusters se pueden usar para identificar patrones o agrupaciones basados en similitudes entre variables ambientales como PM10, NO2, SO2, etc.

Se hizo un análisis de clusters utilizando *k means*, lo cual conecta los puntos basado en la distancia más corta euclidiana sin tomar en cuenta la fecha de los puntos. Al hacer el análisis de esta manera, los resultados, a pesar de parecer legítimos, son inválidos para el fin de nuestro objetivo dado que los clusters creados con los diferentes variables no pertenecían a la misma fecha. Inicialmente se probó el modelo con k=4 lo cual desplegó lo siguiente:

Esta gráfica no es útil para nuestro objetivo dado a que cada *date* y sus valores respectivos pueden estar compuestos en diferentes clusters, aunque explica la proporción de variación total con un valor de 63%. Después de una investigación, se optó por utilizar DTW (Dynamic Time Warping), una librería de R con la cual se pueden identificar patrones o agrupaciones basados en similitudes entre variables ambientales como PM10, NO2, SO2, etc. Con este método se logró determinar qué estaciones tienen comportamientos similares en términos de calidad del aire.

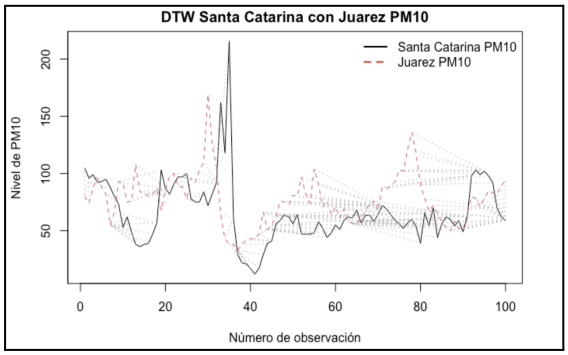


La prueba de Granger se usa para determinar si una serie temporal puede prever otra, verificando relaciones de causalidad temporal. La causalidad de Granger es útil cuando el objetivo es estudiar la relación temporal entre variables (por ejemplo, si las concentraciones

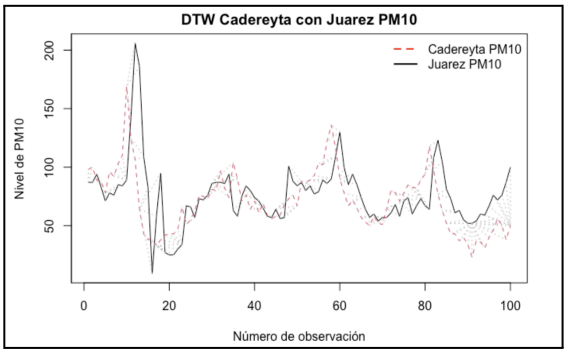
de NO2 en Juárez afectan a Santa Catarina). Esto respalda la hipótesis de que un contaminante puede tener efectos retardados en otras regiones. Se realizaron las siguientes pruebas en base a nuestro objetivo, se encontró que (p-value).

Municipios	**PM10**	**CO**	**SO2**	**NO2**
Juárez causa Santa Catarina	0.01	0.15	0.48	0.85
Cadereyta causa Juarez	0.000001	0.81	0.0000000007	0.91
Cadereyta causa Santa Catarina	0.07	0.85	0.88	0.08

Para la mayoría de los casos, no se encuentra dependencia de los valores pasados de un municipio hacia otra, excepto en el caso de la partícula PM10 y S02. Se rechaza la hipótesis nula en solo dos casos para PM10, donde se establece que el valor pasado de partículas PM10 en Juárez, son significantes para el valor futuro de la misma en Santa Catarina y a la vez el valor pasado de PM10 en Cadereyta causa Santa Catarina. De acuerdo con el análisis, PM10 es una variable clave para el modelo debido a su impacto en la calidad del aire y la salud humana. Se realizó el análisis de DWT para la variable PM10 en los municipios donde se encontró correlación.



DWT Santa Catarina y Juárez



DWT Cadereyta y Juárez

En ambos gráficos se observa como la línea roja punteada (cual representa la partícula PM10 del eje Y) establece la tendencia de altos y bajos, cual la línea sólida negra (cual representa la partícula PM10 del eje X). La línea clara conectando estas dos representa la secuencia en la que una serie de tiempo afecta a la otra. Es decir, cuando se observa que un punto de la línea sólida (time series 1) tiene varias conecciones hacia la línea roja (time series 2) esto significa que la segunda serie de tiempo se está moviendo más lento o con un lag en comparación a la primera. Los movimientos diagonales indican que ambas series de tiempo están moviéndose en conjunto. Para poder ver qué tan similares realmente son las series de tiempo se realiza la prueba similarity score, un valor bajo (0) de este indica que son idénticas, mientras que un valor alto (arriba de 0.5) indica que son diferentes. Al normalizar los scores se obtiene:

Similarity Score (Juárez-Santa Catarina): 0.004752852
Similarity Score (Juárez-Cadereyta): 0.2785171

Ambos valores son extremadamente bajos, lo cual indica una alta correlación entre los municipios.

Relaciones de Dependencia

Adicionalmente a evaluar la relación entre estaciones, se decidió emplear un modelo que permita predecir los valores de un contaminante en cierta región un día antes, para poder anticipar su clasificación con el índice de Aire y Calidad. Para la modelación, se optó por usar un modelo ARIMA. Uno de los supuestos del modelo es que la serie de tiempo debe de ser estacionaria. Para lograr la modelación, se probó el supuesto de si es estacionaria con el test ADF con la siguiente hipótesis

Hipótesis Nula: La serie no es estacionaria.
Hipótesis Alternativa: La serie sí es estacionaria.

	CO	**NO2**	**SO2**	**PM10**
Cadereyta	Estacionaria	No estacionaria	Estacionaria	Estacionaria
Juarez	No estacionaria	No estacionaria	Estacionaria	Estacionaria
Santa Catarina	No estacionaria	Estacionaria	No estacionaria	Estacionaria

Predicciones con ARIMA por Día

Las series no estacionarias fueron diferenciadas una vez para permitir el uso de Arima, se probó con ADF de nuevo y todas las series pasaron la prueba. Para la modelación con ARIMA, se decidió realizar un auto ajuste de parámetros con la función auto.arima() para cada modelo y se obtuvieron los siguientes resultados usando como conjunto de entrenamiento los datos hasta el primero de julio de 2024 y el restante como conjunto de prueba.

Estación	**Contaminante**	**MAE**	**MAPE**
Juarez	CO	0.16138139	22.44372
Juarez	NO2	0.91903443	31.46214
Juarez	PM10	9.56851476	15.47638
Juarez	SO2	1.8816587	86.61093
Santa Catarina	CO	0.06308189	12.85416
Santa Catarina	NO2	2.47336971	16.97157

Santa Catarina	PM10	11.2281346	22.42055
Santa Catarina	SO2	0.46958967	14.50368
Cadereyta	CO	0.20342399	53.65588
Cadereyta	NO2	1.00000003	21.85683
Cadereyta	PM10	11.1698401	18.0801
Cadereyta	SO2	1.67065831	31.66253

Obtuvimos resultados mixtos en la modelación. Para la variable PM10, como se observa en el siguiente gráfico, el ajuste fue acertado para todas las estaciones con un error porcentual promedio menor a 15%. En el caso de Cadereyta se observa bastante similitud y captación de picos en la prueba con los últimos 30 días de registros.

Para otras variables como SO2 en la estación Juárez, se obtuvieron resultados no favorables. Debido a la naturaleza de la variable, sus valores tienden a ser cercanos a cero, lo que un fallo causa que el error porcentual se eleve de manera extrema. El error medio absoluto fue de 1.67 lo que probablemente no sea suficiente para que cambie de nivel en el plan de contingencia ambiental.

Para relacionar este modelo con el análisis de causalidad realizado anteriormente se realizó una comparación de los valores de predicción en Cadereyta contra los valores con Lag 1 de las estaciones de Juárez y Santa Catarina y se obtuvieron las siguientes métricas para los últimos 30 días de la base de datos:

	MAE	**MAPE**
Juárez	13.27856	20.75529%
Santa Catarina	12.90918	27.70493%

El análisis por día resultó en predicciones exitosas, particularmente para la estación de Juarez con un error porcentual promedio de 20% .

Predicciones con ARIMA por Hora

Debido a que las relaciones de Causalidad fueron establecidas en términos de horas, se buscó realizar la misma modelación para la predicción de la siguiente observación para la estación de Cadereyta. Se empleó el mismo modelo ARIMA y se obtuvieron los siguientes resultados para la predicción de PM10 en Cadereyta:

- **MAE**: 12.50491

- **MAPE**: 32.59447%

De la misma manera que el modelo por día, se realizó la comparación de las predicciones contra los datos reales de PM10 en Juárez y Santa Catarina con un ‘Lag’ de una hora. Se obtuvieron los siguientes resultados:

	MAE	**MAPE**
Juárez	28.14458	46.80476%
Santa Catarina	26.08358	72.75841%

Resultados

A continuación, se presentan los principales hallazgos obtenidos a partir del análisis de los datos de calidad del aire recopilados en las estaciones de monitoreo de Cadereyta, Santa Catarina y Juárez. Estos resultados se fundamentan en métodos estadísticos y técnicas de modelado predictivo, así como en la evaluación de patrones temporales y espaciales de los contaminantes atmosféricos monitoreados.

Análisis Estadístico

- **Correlaciones entre contaminantes:**
  Se observaron correlaciones altas entre ciertos contaminantes, como NO y NOX (r = 0.95), lo que sugiere que comparten una fuente común. Por otro lado, contaminantes como SO2 presentaron correlaciones débiles con el resto de las variables, indicando fuentes independientes o procesos distintos de generación.

- **Distribuciones de contaminantes por estación:**
  - **Cadereyta:**
    Se identificaron valores extremos en NO y NOX, reflejando el impacto de fuentes industriales, particularmente la refinería. El PM10 también mostró una alta variabilidad, con episodios ocasionales de concentraciones elevadas.
  - **Santa Catarina:**
    Aunque se observaron niveles moderados para la mayoría de los contaminantes, NOX y PM10 destacaron por sus altos valores en episodios específicos, posiblemente asociados a fuentes locales.
  - **Juárez:**
    Exhibió patrones similares a Santa Catarina, aunque con mayor frecuencia de picos en PM10, lo que refuerza la necesidad de atención en esta región.

- **Medias y varianzas:**
  Las distribuciones de contaminantes mostraron un sesgo a la derecha, indicando que los valores medios están influenciados por episodios extremos de alta contaminación. Esto es particularmente evidente en variables como NO, NOX y PM10 en las tres estaciones.

Modelos Predictivos

- **Predicción con ARIMA:**
  Se implementaron modelos ARIMA para predecir concentraciones de contaminantes clave. Los resultados destacaron:

  Por día:

  - Para PM10 en Cadereyta, los modelos lograron capturar tendencias y picos con un error porcentual promedio menor al 15%.
  - Predicción con Lag en Juárez y Santa Catarina: Error promedio medio en Juarez de 20.75529% y 27.70493% en Santa Catarina

  Por hora:

  - Para PM10 en Cadereyta, los modelos lograron capturar algunas tendencias y picos con un error porcentual promedio menor al 35%.
  - Predicción con Lag en Juárez y Santa Catarina: Error promedio medio en Juarez de 46.80476% y 72.75841% en Santa Catarina
- **Relaciones entre estaciones:**
  El análisis de dependencia temporal mediante pruebas de causalidad de Granger reveló que las concentraciones de PM10 en Cadereyta influyen significativamente en las de Santa Catarina. Además, Juárez mostró dependencia de los valores pasados de PM10 en Santa Catarina, destacando la interacción espacial entre estas regiones.

Identificación de Patrones

El análisis de similitud temporal con técnicas como Dynamic Time Warping (DTW) permitió identificar comportamientos sincronizados entre estaciones, especialmente para PM10. Las similitudes fueron altas, con puntuaciones de similitud bajas (por ejemplo, 0.0047 entre Juárez y Santa Catarina), confirmando patrones comunes en la distribución de partículas.

Visualización Comparativa

Los boxplots generados para cada contaminante permitieron identificar diferencias en la magnitud y frecuencia de los valores extremos entre estaciones. Cadereyta presentó las mayores concentraciones promedio, mientras que Santa Catarina y Juárez mostraron variaciones moderadas, pero con episodios críticos en PM10 y NOX.

Conclusión de los Resultados

Los datos analizados reflejan una dinámica compleja de contaminación atmosférica, influenciada por fuentes industriales y locales. Los resultados obtenidos proporcionan una base sólida para diseñar estrategias de mitigación y prever eventos críticos de contaminación en el área metropolitana de Monterrey.

Discusión y Conclusiones

El reto presentado sobre la calidad del aire en Monterrey proporciona un análisis de las concentraciones de contaminantes atmosféricos en tres estaciones clave: Cadereyta, Santa Catarina y Juárez. A través de métodos estadísticos, modelos predictivos y técnicas de minería de datos, se identificaron patrones temporales y espaciales que destacan la influencia de fuentes industriales y factores locales en los niveles de contaminación.

En particular, las estaciones presentan un sesgo a la derecha en las distribuciones de variables como NO, NOX y PM10, lo que refleja la presencia de valores extremos asociados a episodios críticos de contaminación. Cadereyta, debido a su proximidad a la refinería, reporta consistentemente las mayores concentraciones promedio de CO, NO y NOX, mientras que Santa Catarina y Juárez exhiben niveles significativos de PM10 que influyen de manera significativa la calidad del aire de la zona.

La modelación con ARIMA permitió capturar tendencias y picos de PM10 con éxito moderado, aunque con limitaciones para contaminantes como SO2, debido a su baja variabilidad. Además, el análisis de causalidad de Granger y las pruebas de similitud temporal revelaron relaciones significativas entre estaciones, confirmando la interacción espacial en la distribución de contaminantes. Al tratar de Modelar la causalidad con predicciones, no se logró capturar la relación obtenida a nivel ‘hora’, sin embargo se logró establecer una conexión a nivel ‘día’ entre las estaciones Cadereyta y Juárez.

Cómo siguientes pasos, se propone realizar un análisis detallado para identificar el lag óptimo en las relaciones de causalidad entre estaciones, considerando variaciones temporales que puedan mejorar la precisión de los modelos predictivos y capturar efectos retardados más significativos. Además, se recomienda integrar variables meteorológicas como temperatura, presión atmosférica, dirección y velocidad del viento para evaluar su influencia en la dispersión y acumulación de contaminantes, lo que podría enriquecer los modelos actuales y proporcionar una visión más integral de la dinámica de la contaminación. Por último, es necesario extender el análisis de causalidad a otras estaciones de monitoreo dentro del área metropolitana, con el objetivo de identificar patrones espaciales más amplios y fortalecer las estrategias regionales de gestión ambiental.

Referencias

aire.nl.gob.mx. (n.d.). Gob.Mx. Retrieved November 7, 2024, from http://aire.nl.gob.mx/

de Medio Ambiente y Recursos Naturales, S. (n.d.). Reducir la contaminación atmosférica, tarea de todos. Gob.Mx. Retrieved November 7, 2024, from https://www.gob.mx/semarnat/articulos/reducir-la-contaminacion-atmosferica-tarea-de-todos

¿Qué es el Sistema Integral de Monitoreo Ambiental (SIMA) y cómo mide la calidad del Aire? (2020, July 10). Centro Civitas. https://ccivitas.mx/civitastalks-009/

Abulkhair, A. (2023, 15 junio). Data Imputation demystified | Time Series Data - Ahmed Abulkhair - medium. *Medium*. https://medium.com/@aaabulkhair/data-imputation-demystified-time-series-data-69bc9c798cb7

Link a Bases y Reportes

https://drive.google.com/drive/folders/1-FlZEJ_sa8jN3J8RnEjedmtWbIW5OiFd?usp=drive_link