

Peligros microscópicos: un estudio con estadística multivariada de la contaminación por PM10 en Monterrey

Luis F. Villaseñor, Ernesto I. Borbón, J. de Jesús Gutiérrez, Gerardo Villegas

Docentes: Dra. Cecilia Ramirez, MSE. Ángel Javier Valdez, Mtro. Luis Alonso Hernández

Resumen—El material particulado es una mezcla compleja que permanece suspendida en la atmósfera. Cuando su diámetro aerodinámico es menor o igual a 10 micras se nombra PM10. Altos niveles de PM10 se han relacionado con muertes por causas respiratorias y cardiovasculares. En esta investigación para el Sistema de Monitoreo Integral Ambiental de Monterrey se emplean métodos estadísticos multivariados con los datos de la estación Centro desde el 2017 hasta mediados del 2021 para inquirir en el comportamiento del PM10. Se descubre que hay mayores concentraciones durante el invierno y que aumentaron en 2021. También se encontró que la proporción de registros con PM10 fuera de norma es mayor cuando el viento se dirige al sur y al suroeste. Se ajustaron 2 tipos de regresores para predecir la concentración, por un lado los regresores lineales no lograron cumplir con todos los supuestos y por el otro el regresor random forest obtuvo un R^2 de 0.72. Posteriormente se crearon clasificadores capaces de predecir si el PM10 estará dentro o fuera de norma consiguiendo una exactitud del 91 %. Finalmente se realizó un análisis de conglomerados con 7 componentes obtenidos de un análisis de componentes principales, lo que permitió inferir que hay una asociación de los decrementos de temperatura y velocidad de viento con el aumento de niveles de PM10.

Index Terms—PM10, calidad del aire, estadística multivariada, calendario

I. INTRODUCCIÓN

El material particulado es una mezcla compleja de sustancias en estado líquido o sólido, que permanece suspendida en la atmósfera por períodos variables de tiempo. De acuerdo con su diámetro aerodinámico, éstas pueden clasificarse en menores o iguales a 10 micras (PM10), en menores o iguales a 2.5 micras (PM2.5) y menores o iguales a 0.1 micras (PM0.1). En términos generales, las partículas están formadas por un núcleo de carbono y por compuestos orgánicos e inorgánicos, adheridos a su superficie, y sus concentraciones se miden en microgramos sobre metro cúbico ($\mu\text{g}/\text{m}^3$) [1].

Los efectos de altos niveles de PM10 se han relacionado con exacerbaciones de enfermedades de las vías respiratorias y muertes por causas respiratorias y cardiovasculares [2].

El objetivo de este trabajo es realizar un estudio estadístico multivariado a profundidad de los niveles de PM10 registrados en la estación centro de la base de datos proporcionada por el Sistema Integral de Monitoreo Ambiental (SIMA), que abarca desde inicios de 2017 hasta junio de 2021. Esto permitirá dilucidar las relaciones entre el PM10, otros contaminantes y variables meteorológicas, para así elaborar conclusiones que podrían servir para evaluación de riesgos y toma de decisiones.

Luis F. Villaseñor, Ernesto I. Borbón, J. de Jesús Gutiérrez y Gerardo Villegas pertenecen al Tec de Monterrey Guadalajara, Jal C.P. 45138, Mexico

El resto del trabajo se encuentra organizado de la siguiente manera: Sección II describe más a detalle la problemática, tanto a nivel social como a nivel técnico. Sección III plantea las preguntas detonantes para el análisis de la base de datos. Sección IV expone los pasos realizados y los modelos utilizados para el estudio de los datos. Por último, la sección VI presenta los resultados y conclusiones, al igual que recomendaciones a futuro.

II. DESCRIPCIÓN DE LA PROBLEMÁTICA

Las partículas suspendidas menores a 10 micrómetros de diámetro son las más dañinas para la salud debido a que pueden penetrar y alojarse en el interior profundo de los pulmones, aumentando el riesgo de desarrollar patologías como cardiopatías y neumopatías, al igual que cáncer de pulmón [3].

Además se ha encontrado que hay una relación entre la exposición de altas concentraciones de pequeñas partículas y el aumento de la mortalidad diaria y a largo plazo. No obstante no se ha podido identificar ningún umbral por debajo del cual no se hayan observado daños para la salud [3].

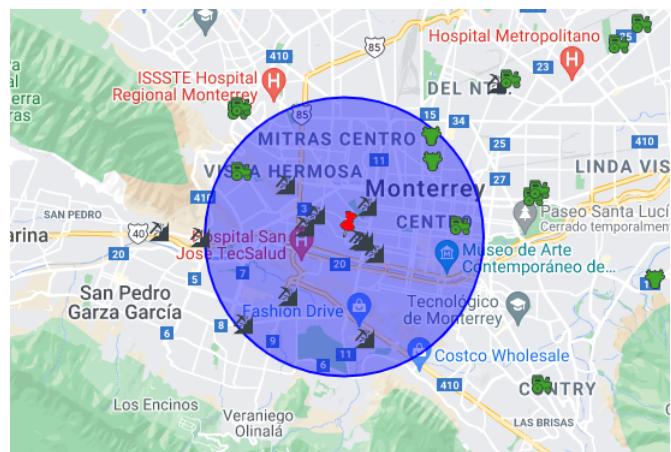


Figura 1. Ubicación de la estación. El círculo representa el rango de los sensores (4km). Los íconos negros indican los puntos donde se llevan a cabo actividades de extracción de petróleo, minería y gas. Los íconos con tractor indican actividades agrícolas, y los íconos de vaca representan actividades ganaderas.

Por lo anterior, organismos tanto gubernamentales como no gubernamentales han propuesto iniciativas y acciones para monitorear y reducir la contaminación del aire. Por ejemplo, en México la NORMA Oficial Mexicana NOM-025-SSA1-2014 de Salud ambiental establece los valores límite permisibles para la concentración de partículas suspendidas PM10 y PM2.5

en el aire ambiente y criterios para su evaluación. En esta norma se especifica que el valor promedio límite en 24 horas para el PM10 es 75 $\mu\text{g}/\text{m}^3$, mientras que el anual es de 40 $\mu\text{g}/\text{m}^3$ [1].

Con los datos que recaba el SIMA, se puede vigilar el cumplimiento de las normas. La estación cuyos datos se utilizarán en este análisis se encuentra en las instalaciones de Servicios de Agua y Drenaje de Monterrey en el área del Obispado. La ubicación de esta estación pretende monitorear la contaminación de fuentes vehiculares e industriales y permite evaluar los efectos de la mezcla de contaminantes en una zona residencial del centro del Área Metropolitana de Monterrey [4]. En la Fig. 1 se visualizan la localización de la estación centro y las actividades económicas que circundan según el Directorio Estadístico Nacional de Unidades Económicas [5]. Se puede observar una especial concentración de actividades de extracción de petróleo, minería y gas.

Se cuentan con más de 39,000 observaciones, sin embargo no se tienen todas las mediciones de todas las variables debido a fallas en los sensores o circunstancias extraordinarias. Esto supone otro reto, tanto para el preprocesamiento de los datos, como para la aplicación de los métodos multivariados.

III. PREGUNTAS DE INVESTIGACIÓN

Para guiar el análisis de los datos y del comportamiento del PM10 se plantearon las siguientes preguntas:

1. ¿Existe una variación temporal del PM10?
2. ¿Cómo podríamos predecir la concentración de PM10 considerando otros contaminantes y variables meteorológicas?
3. ¿Se han reducido los niveles de contaminación por PM10?
4. ¿Con qué exactitud se podría categorizar un día dentro o fuera de la norma en cuanto a concentraciones de PM10 teniendo el resto de las variables?
5. ¿Qué variables están asociadas a altas concentraciones de PM10?

IV. METODOLOGÍA UTILIZADA

IV-A. Limpieza y preparación de los datos

IV-A1. Eliminación de datos no requeridos o inválidos: Inicialmente se tenían 8 conjuntos de datos de diferentes contaminantes y 7 conjuntos de datos con variables meteorológicas. Cada conjunto de datos consistía en la medición de una variable en las 14 estaciones del SIMA en la zona metropolitana de Monterrey. Cuentan con un registro por hora desde el 1º de enero del 2017 a las 00 horas, hasta el 30 de junio del 2021 a las 23 horas, resultando en 39,394 registros. Se extrajeron únicamente aquellos datos de la estación Centro y se eliminaron los datos que tuvieran “banderas” inválidas. Las banderas pueden resultar inválidas por fallas eléctricas, calibración de los sensores, valores atípicos, entre otras razones similares.

IV-A2. Fusión de los conjuntos de datos: Se fusionaron los 15 conjuntos de datos originales que contenían una sola variable meteorológica o de un contaminante en un solo conjunto de datos con todas las variables.

IV-A3. Imputación de valores faltantes: Con el fin de no perder los registros con algunos valores faltantes, se realizó una imputación para todas las variables. Ésta consiste en repetir el siguiente valor válido en el conjunto de datos, siempre y cuando éste haya ocurrido como máximo 48 horas después (rellenado hacia atrás). Adicionalmente se realizó otra imputación que toma el último valor válido con la misma regla de las 48 horas (rellenado hacia adelante). De este modo se pueden llegar a llenar hasta 96 horas de datos faltantes.

IV-A4. Eliminación de registros con valores faltantes: Una vez realizada la imputación de datos, se procedió a eliminar los registros con valores faltantes para que no interfirieran con los modelos de regresión y clasificación. Esto es porque los algoritmos fallan al encontrar datos nulos.

IV-A5. Creación de variables auxiliares y elaboración del conjunto con promedios de concentración PM10 por día: Se agregaron variables que facilitaron el análisis como el día, el mes, la estación del año, el año y la fecha sin la hora. Con estas variables temporales se pudo crear un nuevo conjunto de datos que agrupa a los días y calcula su promedio de concentración de PM10. En este conjunto nuevo se creó una variable categórica llamada “Peligro” que indica si el riesgo por PM10 es bajo, moderado o alto, inspirada en la NOM-172-SEMARNAT-2019 [6]. Los rangos para cada categoría se despliegan en la tabla I.

Tabla I
RANGOS DE RIESGO POR CONCENTRACIÓN DE PM10

Riesgo	Intervalo de PM10 ($\mu\text{g}/\text{m}^3$)
Bajo	<50
Moderado	50 - 75
Alto	>75

También se creó otra variable de nombre “PM10_pel”, que indica si el riesgo de PM10 está dentro o fuera de los valores límites establecidos según la NOM-025-SSA1-2014. Los rangos para cada categoría se despliegan en la tabla II.

Tabla II
RANGOS DE LOS VALORES LÍMITES POR CONCENTRACIÓN DE PM10

Riesgo	Intervalo de PM10 ($\mu\text{g}/\text{m}^3$)
Dentro de los valores límites	<75
Fuera de los valores límites	>75

IV-B. Análisis exploratorio

IV-B1. Visualizaciones: Se obtuvieron los promedios diarios de concentraciones de PM10 y se realizaron gráficas de caja y un calendario en donde se muestran si las concentraciones fueron bajas, moderadas o altas. También se realizó una gráfica circular con la dirección del viento y las concentraciones de PM10 correspondientes utilizando la librería windrose.

IV-B2. Correlaciones: Se computaron los coeficientes de correlación de Pearson para determinar si hay relaciones lineales del PM10 con las demás variables.

IV-C. Análisis de Regresión

Con el fin de obtener resultados más precisos, se dividió el conjunto de datos original en 4 subconjuntos nuevos, uno por

cada estación del año. La variable dependiente de los modelos sería la concentración de PM10, y las variables independientes serían las concentraciones del resto de contaminantes (PM2.5, CO, NO₂, NO, NO_x, SO₂), el día, el mes y las mediciones de la temperatura (TOUT), la humedad relativa (RH), la radiación solar (SR), la precipitación (RAINF), la presión atmosférica (PRS), la velocidad del viento (WSR) y de la dirección del viento (WDR).

IV-C1. Regresión lineal múltiple: Por cada estación se comenzó estimando un modelo OLS con la librería statsmodels [7] utilizando todas las variables. Se revisaron los coeficientes estimados y aquellos con un valor p menor a 0.05 eran catalogados como insignificantes y se ajustaba un nuevo modelo sin esas variables. Después se revisaba el factor de inflación de la varianza (FIV) de cada variable para evaluar multicolinealidad, y aquellas con un FIV mayor a 5 fueron removidas. Lo anterior se realizó por iteraciones hasta tener un modelo con solo coeficientes significativos y todos los FIV menores a 5. Una vez logrado esto se evaluaron visualmente los demás supuestos de los modelos lineales:

- Normalidad de los residuos.
- Homocedasticidad.
- Independencia de los errores.

IV-C2. Regresor Random Forest: Como alternativa al regresor lineal múltiple, se realizó un regresor Random Forest. Además, se utilizó el método de búsqueda aleatoria para la optimización de parámetros y una regresión lasso para la selección de características.

IV-D. Análisis de clasificación

Se realizaron ANOVAs de un solo factor entre la variable Peligro y las demás variables de la base de datos con nivel de significancia de 0.05. Lo mismo para la variable PM10_pel. Las variables que resultaron significativas fueron utilizadas para un análisis discriminante y para un clasificador de random forests. Para revisar la efectividad de cada modelo se imprimió una matriz de confusión.

IV-E. Reducción de dimensionalidad

Con el fin de simplificar el problema y el número de variables, es necesario aplicar algún algoritmo de reducción de dimensionalidad. Esto permitirá aplicar otros tipos de algoritmos para los cuales su efectividad se ve asistida por tener menos variables.

IV-E1. Análisis de componentes principales: Se usó la función PCA de la librería de Scikit Learn, luego se definió un número de componentes que lograran explicar suficiente variabilidad y después se interpretó cada uno de ellos.

IV-E2. Análisis factorial: Se utilizó la librería FactorAnalyzer para realizar la prueba de esfericidad de Bartlett y calcular el índice Kaiser-Meyer-Olkin y determinar la viabilidad de un análisis factorial.

IV-F. Análisis de conglomerados

Para definir el número de conglomerados a utilizar se empleó el método de codo. Posteriormente se aplicó la función de

KMeans de la librería de Scikit Learn. Al obtener los grupos, se inquirió en cuál se contienen las mayores concentraciones de PM10 en promedio y se caracterizó cada uno de ellos en función de los componentes provistos por el PCA.

V. RESULTADOS

V-A. Limpieza y preparación de datos

El conjunto de datos que resultó tras la fusión descrita en IV-A2 tenía un promedio de 9,640 datos faltantes en las variables de los contaminantes y 1,469 para las variables meteorológicas. En cuanto a los datos faltantes de las variables de contaminantes, el que más tuvo fue SO₂ con 18,908 y el que menos tuvo fue PM10 con 1,634. Antes de proceder a eliminar los valores faltantes, se aplicaron los métodos de imputación de datos descritos en la apartado IV-A3. De este modo se lograron imputar un promedio de 2347 datos por variable y la variable PM10 resultó sin datos faltantes. Posteriormente el conjunto de datos que resultó tras la eliminación de registros con datos faltantes (véase IV-A4) fue de 20,154 registros.

V-B. Análisis exploratorio

V-B1. Visualizaciones: Una vez preparado el conjunto de datos, se lograron identificar ciertas tendencias a partir de las visualizaciones creadas. Como se observa en la Fig. 2, 2018 y 2019 presentaron una disminución en la concentración de PM10 respecto al 2017, pero es una tendencia que se pierde al llegar el 2020, que presenta valores similares al 2017. A su vez, el 2019 es el año que presenta la mayor variabilidad. Nótese que solo hay registros de la primera mitad del 2021, y con este periodo las concentraciones promedio son bastante elevadas en comparación con los promedios diarios de años anteriores.

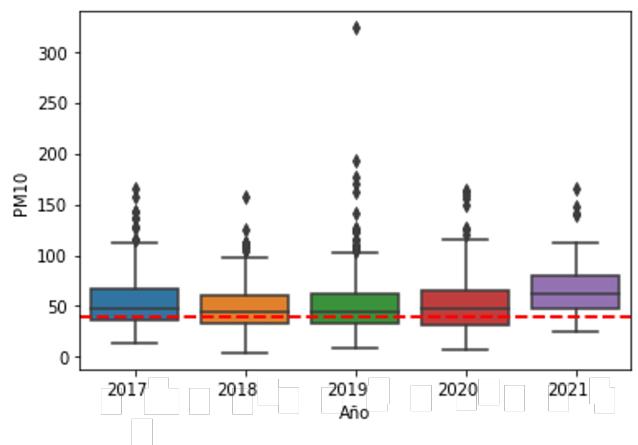


Figura 2. Gráficas de caja de concentraciones PM10 promedio diarias por año (medio año en caso de 2021). La línea roja punteada indica una concentración de 40 µg/m³.

Analizando las concentraciones de PM10 por mes en la Fig. 3, es posible observar que los meses más contaminados son los de invierno y la concentración disminuye conforme se acerca el verano, a excepción de julio en donde hay un repunte.

Estas mismas tendencias se pueden apreciar en la Fig. 4 en donde destaca que en lo va del 2021 la mayoría de días

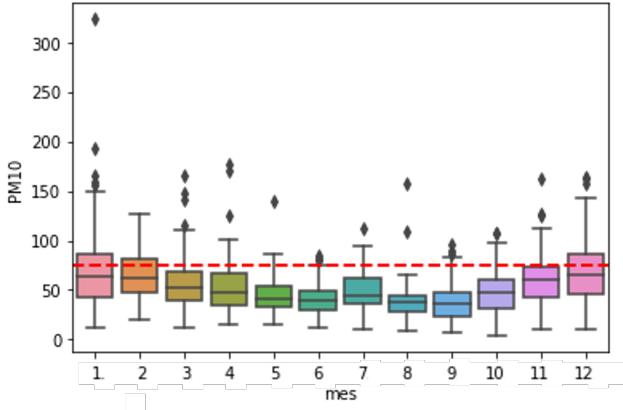


Figura 3. Gráficas de caja de concentración promedio diaria de PM10 por mes. La línea roja punteada indica el valor límite diario de $75 \mu\text{g}/\text{m}^3$.

tienen una concentración promedio de PM10 moderada o alta. Más aún, los días fuera de norma hasta junio superaron en número a todo el año de 2017, 2018 y 2020. También es fácil de ver como fuera del 2021, los inviernos son los que más días tienen con concentraciones moderadas y altas de PM10. Finalmente se puede observar que el mes de julio es el más contaminado a mediados de año.

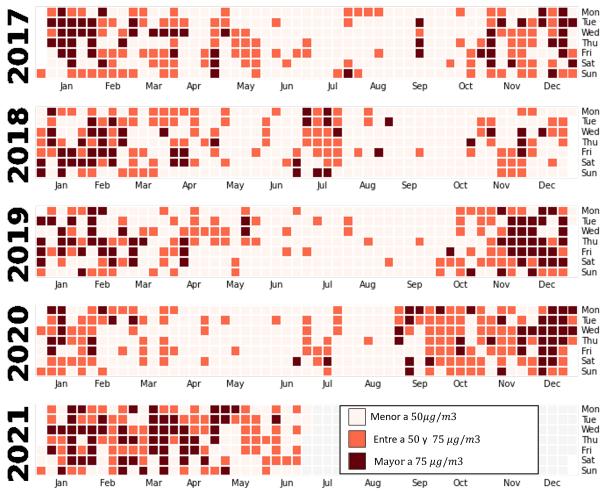


Figura 4. Calendario de niveles de riesgo de las concentraciones promedio diarias de PM10.

Adicionalmente, al analizar las concentraciones de PM10 por hora en la Figura 5, se puede observar que aunque la concentración de PM10 no varía mucho durante el día, las mayores concentraciones se alcanzan entre las 11:00 y las 13:00 horas. Esto puede sugerir que los picos de PM10 son causados por variables meteorológicas y no antropogénicas, dado que de las 11:00 a las 13:00 horas no son horas particulares como la entrada o salida del trabajo o la escuela, esto en cuestión del tráfico. Se tendría que revisar si la actividad industrial aumenta en dichas horas.

Para terminar el análisis exploratorio se analizaron los registros con PM10 dentro y fuera de norma dependiendo de la dirección del viento. Se descubrió a través de una prueba de hipótesis que la proporción de registros con PM10 fuera

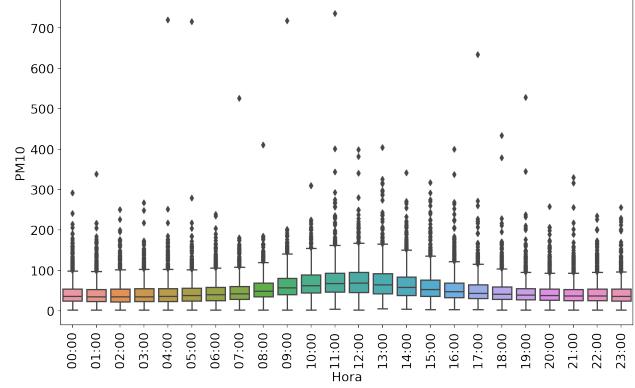


Figura 5. Gráficas de caja de concentración de PM10 por hora del día.

de norma es mayor cuando el viento se dirige hacia el sur o suroeste (170° - 235°). Esto se refleja en que la Figura 8 que tiene los registros con PM10 fuera de norma tiene un aumento considerable en las direcciones sur y suroeste con respecto a la Fig. 7.

V-B2. Correlaciones: La Fig. 6 presenta las correlaciones de Pearson de cada variable con el PM10. Como era de esperarse, la correlación con PM2.5 es la que tiene el coeficiente más alto (0.689) e indica que hay una correlación lineal positiva moderada. La correlación con el resto de variables no es buena y no parece revelarse algún vínculo entre una variable meteorológica con el PM10 con este análisis. Esto augura que el modelo de regresión lineal no será muy exitoso.

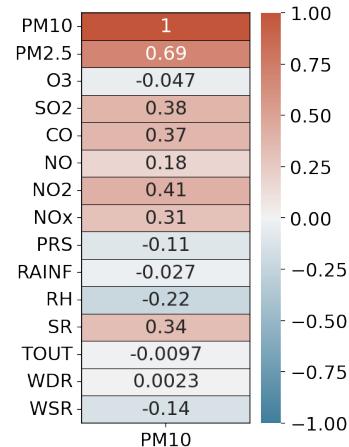


Figura 6. Coeficientes de correlación de Pearson de cada variable con el PM10.

V-C. Análisis de regresión

V-C1. Regresión lineal múltiple: Los 4 modelos estimados cumplen con la no multicolinealidad, la normalidad de los residuos y la independencia de los errores, sin embargo ninguno cumple con la homocedasticidad (que la variabilidad de los residuos sea constante). Esto hace que el modelo no sea tan fiable. Con el propósito de enmendarlo se transformó logarítmicamente la variable dependiente PM10.

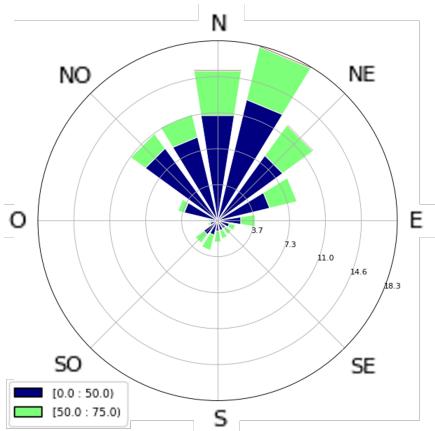


Figura 7. Proporción de días con PM10 dentro de los valores límites por dirección.

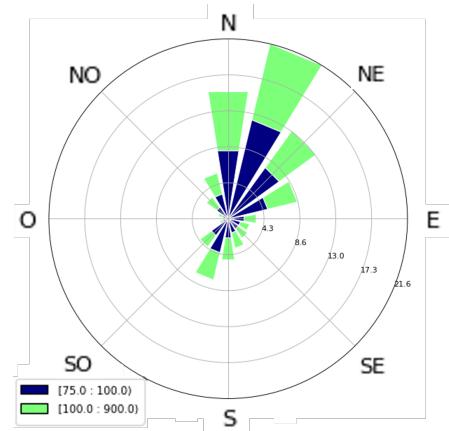


Figura 8. Proporción de días con PM10 fuera de los valores límites por dirección.

En la tabla III se muestran los coeficientes de determinación ajustados de cada modelo. Entre más cerca esté de 1, más explica el modelo la variabilidad del fenómeno. Los modelos de primavera e invierno mejoraron su desempeño con la transformación de variable, pero sus residuos siguen siendo heterocedásticos.

Tabla III
VALORES R² AJUSTADOS DE LOS MODELOS POR CADA ESTACIÓN Y ESTADO DE LA VARIABLE PM10

Variable	Otoño	Invierno	Primavera	Verano
Transformada	0.554	0.530	0.422	0.269
Original	0.573	0.498	0.375	0.413

Como se puede ver, los datos más explicables son los del otoño y el invierno, y los menos explicables son los del verano. Esto se puede deber a las condiciones meteorológicas. Al aplicar pruebas de hipótesis de diferencia de medias en las variables climáticas del otoño e invierno, se encontró con un nivel de significancia de 0.05 que son distintas las medias de todas las variables meteorológicas a excepción de la precipitación. En verano en promedio hay menos presión atmosférica, menos humedad relativa, más radiación solar, más temperatura y velocidad del viento más alta que en otoño.

En el anexo se puede consultar el resumen de modelo con sus respectivas visualizaciones de revisión de supuestos y tablas de factores de inflación de la varianza.

En la tabla IV se despliegan los coeficientes del mejor modelo para cada estación. Los coeficientes revelan la importancia de las variables para entender el fenómeno y es importante notar que los coeficientes más altos en todos los modelos corresponden a la radiación solar. Las variables que menos ayudan a estimar los datos son el NO y el NOx, pues son variables que fueron eliminadas en al menos 3 de los 4 modelos. Esta remoción se explica por su alta colinealidad con NO2.

V-C2. Regresor Random Forest: En la Fig. 9 se graficó el valor absoluto de los diferentes coeficientes en la regresión lasso, que se pueden interpretar como la importancia de cada variable. De modo que las variables más importantes para

Tabla IV
COEFICIENTES DE LOS MODELOS LINEALES

	Otoño	Invierno	Primavera	Verano
Constante	56.72	10.72	21.35	1303.51
PM2.5	0.828	0.012	0.015	0.94
O3	-	-0.005	-0.007	0.121
SO2	0.524	0.031	0.033	0.622
CO	13.79	0.071	0.031	4.82
NO	-	0.004	-	-
NO2	2.71	0.012	-	1.83
NOx	-	-	-	-
PRS	-	-0.01	-0.024	-
RAINF	1.279e-14	-9.891e-17	5.778e-12	-9.636e-11
RH	-0.213	-0.007	-0.006	0.119
SR	-	0.588	0.588	19.19
TOUT	-	0.016	0.012	-0.5654
WDR	-	0.0003	-0.0006	-
WSR	2.02	0.0109	0.011	-0.611
Día	-0.262	-	-	-0.2863
Mes	-6.28	-	-0.07	2.045

predecir las concentraciones de PM10 son NO2, CO, SO2 y PM2.5.

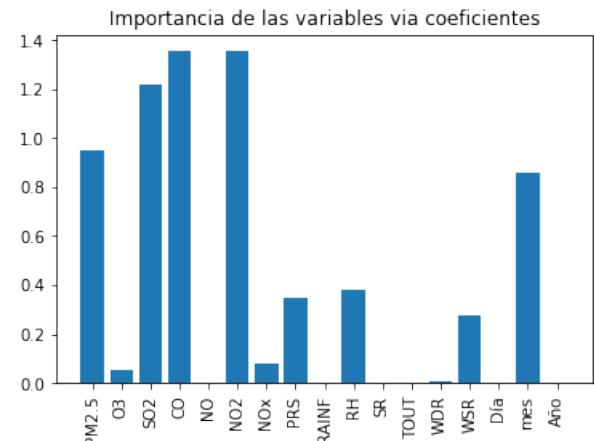


Figura 9. Importancia de cada variable para predecir PM10 según la regresión lasso.

Los parámetros que optimizan al modelo fueron 252 esti-

madores, 5 muestras mínimo para dividir un nodo interno, 1 muestra mínimo por nodo hoja, la raíz cuadrada para determinar el número de características al buscar la mejor división, 94 como máxima profundidad y se utiliza todo el conjunto de datos para armar cada árbol de decisión. Posteriormente se realizaron dos modelos con los parámetros mencionados. Uno con las características más importantes de acuerdo con la regresión lasso presentes en la figura 9 (PM2.5, SO2, CO, NO2, PRS, RH, WSR y mes), y otro con todas las características. Para evaluar los modelos se utilizó el método de *Repeated K-Fold* con 10 divisiones, 3 repeticiones y R^2 como puntuación. El modelo con 18 características obtuvo una puntuación de 0.75 mientras que el de 8 características obtuvo 0.72. Se optó por utilizar el segundo modelo ya que se pierde muy poca explicación de la variabilidad y utiliza muchas menos variables.

V-D. Análisis de clasificación

V-D1. ANOVAs: Tras los ANOVAs las variables significativas fueron PM2.5, SO2, CO, NO, NO2, NOx, PRS, RH, SR, TOUT, WDR, WSR, dia, mes y año; tanto como para el factor Peligro como para PM10_pel.

V-D2. Análisis Discriminante Lineal- “Peligro”: En la figura 10 se muestra la exactitud del modelo, que en este caso es de .68. El grupo más difícil de predecir es el grupo moderado (1) y la precisión de los grupos es de 0.70 para el grupo 1, 0.51 para el grupo 2 y 0.71 para el grupo 3.



Figura 10. Matriz de confusión con LDA empleando la variable Peligro

V-D3. Análisis Discriminante Lineal- “PM10_pel”: Como se puede deducir de la Fig. 11 el modelo empleando la variable PM10_pel es mejor teniendo una exactitud de .86 y precisión de 0.75. Es decir, es más efectivo solo identificar si un promedio diario de concentración está fuera de norma o no lo está.

V-D4. Clasificador de Random Forests (RF) con “PM10_pel”: Este modelo tuvo una exactitud de 0.91 y 0.84 de precisión (véase Fig. 12). Exitosamente logra identificar correctamente 194 datos más que el análisis discriminante.

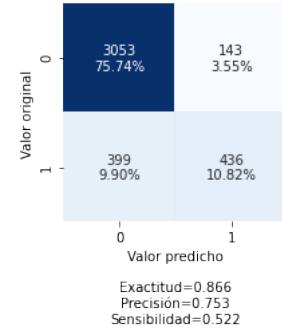


Figura 11. Matriz de confusión con LDA empleando la variable PM10_pel.

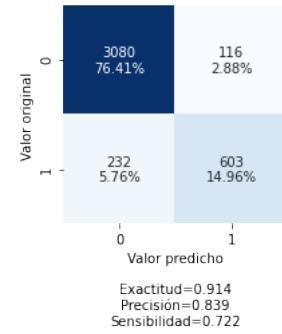


Figura 12. Matriz de confusión con RF empleando la variable PM10_pel.

V-E. Reducción de dimensionalidad

V-E1. Análisis de componentes principales: Una vez hecho el análisis de componentes principales con 10 componentes para observar su comportamiento, se decidió por solo trabajar con 7 componentes, puesto que con este número se describía más del 0.79 de la variabilidad de los datos. En la Tabla V se muestran los primeros tres componentes, los cuales son los más valiosos para el estudio.

Tabla V
ANÁLISIS DE COMPONENTES PRINCIPALES

	Componente Principal 1	Componente Principal 2	Componente Principal 3
PM10	0.20484	0.370791	-0.344478
PM2.5	0.268558	0.296434	-0.322469
O3	-0.375564	0.228771	-0.157696
SO2	0.118824	0.283595	-0.183594
CO	0.288445	-0.013913	-0.380146
NO	0.14331	0.324152	0.560948
NO2	0.342528	0.266421	0.0173
NOx	0.247769	0.370624	0.478151
PRS	0.179061	-0.222218	0.043194
RAINF	-0.003555	0.004685	0.028078
RH	0.251433	-0.328786	0.15409
SR	-0.173131	0.284081	-0.026762
TOUT	-0.385322	0.271853	0.011576
WDR	0.165434	-0.031504	0.002979
WSR	-0.38365	0.123155	0.058447
Ratio	0.228269	0.188537	0.111399

El componente principal 1 se interpreta como Volatilidad, el 2 como PM10 y el 3 como Ausencia de Contaminantes Relacionados.

V-E2. *Análisis factorial:* Se determinó que es inadecuado realizar un análisis factorial con el conjunto de datos ya que, a pesar de haber pasado la prueba de esfericidad de Bartlett, se obtuvo un índice Kaiser-Meyer-Olkin (KMO) de 0.58, el cual es considerado bajo.

V-F. Análisis de conglomerados

No se observaba de manera definitiva el número de conglomerados óptimo al usar el método del codo, por lo que de manera arbitraria se optó por utilizar 6 conglomerados. Se extrajeron los centroides de cada grupo y se computó el valor promedio de PM10 por cada conglomerado, con lo cual se hizo la gráfica de burbuja de la 13 que representa la posición y la comparación de promedios por grupo entre los componentes principales 1 (Volatilidad) y 3 (Ausencia de contaminantes relacionados). No se usó el componente PM10 porque sería redundante.

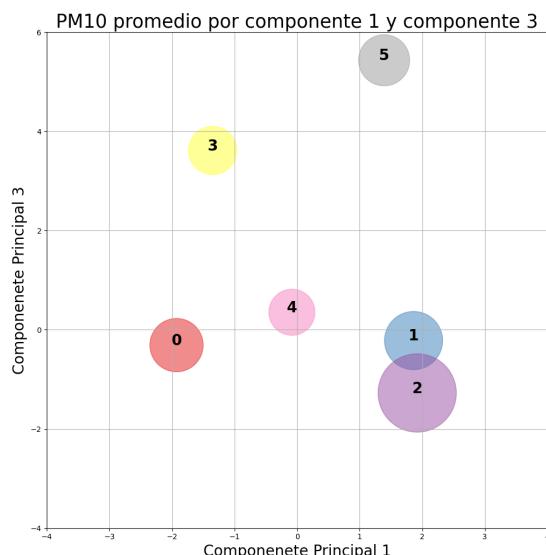


Figura 13. Gráfica de burbujas en donde se muestran los centroides de los conglomerados y el tamaño es el valor promedio de PM10

Es visible la influencia del componente principal 1 en la cantidad de PM10 medida, ya que entre mayor es su puntaje,^[3] mayor es el promedio de concentraciones de PM10. Esto asocia a bajas temperaturas y bajas velocidades de viento con^[4] mayores concentraciones de PM10.

En cuanto al componente principal 3 se puede observar que^[5] entre más abajo esté el centroide, también crecerá el promedio de PM10. Esto es coherente, puesto que los coeficientes de CO y PM2.5 son negativos en dicho componente y son dos variables que ya han probado estar relacionadas positivamente con el PM10 (véase Fig. 6). Por el otro lado, este análisis también deja ver que concentraciones elevadas de NO y NOx no vienen acompañadas de altas concentraciones de PM10.

VI. CONCLUSIONES

Se identificaron dos patrones temporales particulares. El primero es que la concentración de PM10 aumenta en invierno.

Con una mirada más profunda, a nivel horario el aumento de PM10 se da a partir de las 9 de la mañana tiene un pico a las 12 y disminuye de nuevo a las 3.

Se elaboró un calendario que permite visualizar los días fuera de norma en los últimos 4 años, y a través de él se concluye que los niveles de contaminación por PM10 han aumentado. Tanto así que los primeros 6 meses de 2021 superan en días que exceden los valores límites de la norma a los años 2017, 2018 y 2020.

Se logró observar que a través del año los niveles de PM10 tienden a comportarse de manera distinta, por lo que separar el análisis por estación del año resultó ser la mejor opción para obtener regresiones más precisas.

Se generaron varios modelos de regresión lineal que explican del 26% al 55.5% de la variabilidad de los datos dependiendo de la estación del año. Se comprobó que estas variaciones se deben a sus diferencias en las variables meteorológicas. No obstante estos modelos no cumplen con todos los supuestos de los regresores lineales y por lo tanto no son un acercamiento adecuado para este problema. Por otro lado, también se propuso un modelo de Random Forest que explica el 72 % de la variabilidad utilizando las variables PM2.5, SO2, CO, NO2, PRS, RH, WSR y mes.

El análisis de clasificación consiguió ser útil para predecir si un día está fuera o dentro de los límites establecidos, logrando obtener un modelo usando un clasificador de Random Forests y la variable PM10_pel con una exactitud del 91 %.

Se efectuó un análisis de conglomerados con 7 componentes brindados por el PCA. Esto permitió encontrar relaciones entre las variables y el PM10, a saber, que la disminución de temperatura y velocidad de viento están asociadas con el incremento de la concentración del PM10.

Para finalizar, futuros trabajos podrían utilizar un acercamiento desde las series de tiempo para profundizar en el comportamiento de las concentraciones de PM10.

REFERENCIAS

- [1] “Norma oficial mexicana nom-025-ssa1-2014, salud ambiental.” [Online]. Available: http://www.dof.gob.mx/nota_detalle.php?codigo=5357042&fecha=20/08/2014
- [2] W. M. Kenneth Donaldson, “Potential mechanisms of adverse pulmonary and cardiovascular effects of particulate air pollution (PM10),” 2001. [Online]. Available: <https://doi.org/10.1078/1438-4639-00059>
- [3] “Efectos a la salud por la contaminación del aire ambiente.” [Online]. Available: <https://www.gob.mx/cofepris/acciones-y-programas/3-efectos-a-la-salud-por-la-contaminacion-del-aire-ambiente>
- [4] INECC, “MONITOREO DE LA CALIDAD DEL AIRE EN EL ÁREA METROPOLITANA DE MONTERREY,” 2007. [Online]. Available: <http://www2.inecc.gob.mx/publicaciones2/libros/234/cap3.html>
- [5] I. N. de Estadística y Geografía (INEGI), “Directorio estadístico nacional de unidades económicas. denue.” [Online]. Available: <https://www.inegi.org.mx/app/mapa/denue/default.aspx>
- [6] “Norma oficial mexicana nom-172-semarnat-2014, salud ambiental.”
- [7] S. Seabold and J. Perktold, “statsmodels: Econometric and statistical modeling with python,” in *9th Python in Science Conference*, 2010.