ANÁLISIS ESTADÍSTICO DE OZONO A NIVEL DEL SUELO

Ph.D. Héctor Quevedo Urías[1], Ph.D. Humberto García[2], Ph.D. Jorge Salas Plata[1], MIA Angelina Domínguez Chicas[1], Ing. Víctor H. Esquivel Ceballos[1]

RESUMEN

Se aplicaron métodos estadísticos para modelar las concentraciones de ozono (O_3) troposférico observadas en el Parque Chamizal, una estación de muestreo localizada en la línea divisoria entre El Paso, Texas, EU y Ciudad Juárez, Chihuahua, México. El estudio consistió en modelar las concentraciones de O_3 (tomada como la variable dependiente) y la temperatura del punto de rocío, óxido de nitrógeno (NO), dióxido de nitrógeno (NO_2), temperatura ambiental, humedad relativa, radiación solar, intensidad del viento, intensidad del viento resultante, dirección del viento resultante y ráfagas máximas (tomadas como las variables independientes). La metodología usada en este estudio estadístico consistió en ajustar un modelo de regresión lineal múltiple y un modelo de regresión polinomial, a las variaciones espacio-temporales de 1-hora de O_3, para el periodo 1999-2003 (cerca de 470,000 casos). Para la selección del mejor modelo candidato, se depuraron los datos por medio de eliminar los casos atípicos extremos. Otrosí, se aplicaron procedimientos de regresión usando las funciones de Stepwise y Best-subset del programa Minitab, con el objeto de obtener el modelo candidato más refinado. Además, cada modelo de regresión se evaluó usando diagnósticos objetivistas, como el coeficiente de determinación múltiple (R^2), el error estándar de lo estimado (s), errores de predicción (residuales PRESS) y la estadística Cp de Mallows. Este diagnóstico se complementó usando gráficos subjetivistas. El mejor modelo estadístico mostró valores de 70.0%, 10.1, $4.4x10^6$, y 9.0 para R^2, s, PRESS, y la estadística Cp, respectivamente. Finalmente, este estudio incluyó el promedio aritmético y las desviaciones estándar de los valores de O_3 y de las variables independientes.

INTRODUCCIÓN

La razón que motivó este estudio estadístico fue debido a que la cuenca atmosférica de El Paso y Ciudad Juárez se caracteriza por frecuentes eventos de altas concentraciones de ozono a nivel del suelo. Esta región también se caracteriza por situaciones excepcionales meteorológicas, como altas temperaturas y altos niveles de radiación solar (especialmente durante el verano). Esta área también se caracteriza por altas concentraciones de óxidos de nitrógeno y compuestos orgánicos volátiles emitidos por vehículos y actividades industriales.[12]

Al presente, no existen estudios estadísticos de esta índole en el área de Ciudad Juárez. Por lo tanto, debido a esta situación tan desafiante, la presente investigación exploró la posibilidad de desarrollar un modelo de ozono basado un análisis matemático heurístico simple. Con esto, la idea principal era la de estudiar las concentraciones del ozono superficial, particularmente, aquellas concentraciones que estén violando el límite de 1-hora establecido por la Agencia de Protección del Medio Ambiente de Estados Unidos (EPA) y por su contraparte, la Procuraduría de Protección al Ambiente (PROFEPA), para proteger la salud pública. En este respecto, es conveniente mencionar que, ambas agencias están de acuerdo en establecer el promedio aritmético de ozono de 1-hora en 0.12 ppm (120 ppb).

Es bien sabido que los contaminantes primarios atmosféricos emitidos por fuentes estacionarias y móviles (autos, camiones, ferrocarriles, aeroplanos, plantas eléctricas, fábricas, calentadores domésticos, etc.) producen monóxido de nitrógeno (NO) y, hasta cierto punto, compuestos orgánicos volátiles (VOC). En turno, estos gases atmosféricos, cuando son accionados por las condiciones meteorológicas apropiadas, producen el ozono troposférico. Por ejemplo, cuando las fuentes de combustión interna emiten NO, este compuesto químico, a través de la interacción de calor y luz solar, propicia que el NO se oxide rápidamente a dióxido de nitrógeno (NO_2). Enseguida, a través de la acción del calor y de la luz solar, el NO_2 se disocia a O_3. De hecho, la contaminación troposferica de ozono ocurre principalmente durante el medio día en tiempo de verano. De acuerdo a Brown et al. (2000), las reacciones típicas atmosféricas químicas que generan el ozono a nivel del suelo son:

[1] Instituto de Ingeniería y Tecnología, Universidad Autónoma de Ciudad Juárez, Ave del Charro 610 N. Edificio E. Ciudad Juárez, Chihuahua 32310, México.
[2] Instituto Tecnológico y de Estudios Superiores de Monterrey (Campus Ciudad Juárez).

$$NO + VOCs \rightarrow NO_2 \quad (1)$$

$$NO_2 \xrightarrow{\text{luz solar}} NO + O \quad (2)$$

$$O + O_2 \rightarrow O_3 \quad (3)$$

Por otro lado, Davis et al. (1998) desarrollaron modelos estadísticos para predecir el ozono en áreas urbanas basadas en los clásicos modelos de los cuadrados mínimos. Similarmente, otros investigadores (Libiseller et al. 2003) contendieron que la normalización de modelos regionales puede ser superior a los modelos que operan en áreas individuales debido a que la formación de ozono es un proceso muy frecuente y complejo que circunda un área ampliamente distribuida. Además, Hubbard et al. (1998) desarrollaron un modelo de regresión con 10 variables independientes que incluyeron temperatura ambiental, descripción celeste, punto de rocío, humedad relativa, intensidad del viento, temperaturas mínimas y máximas diarias y cantidad de precipitación. Más adelante, otros investigadores (Eder et al. 1993) hicieron estudios para caracterizar la variabilidad espacio-temporal de las concentraciones de ozono en áreas no urbanas en el este de Estados Unidos. Este estudio se hizo paras explicar la relación entre las emisiones de ozono y los patrones meteorológicos.

De esta manera, los objetivos de esta investigación fueron relacionados con la aplicación sistemática de la selección de los mejores análisis de regresiones estadísticas aplicadas al ozono troposferico, datos químicos y datos meteorológicos. Esto se hizo usando un razonamiento estadístico muy profundo, para seleccionar el mejor modelo de regresión. A pesar de que los datos de los promedios de 1-hora de las concentraciones del ozono superficial están correlacionadas en serie, la aplicación de los análisis clásicos de regresión son encomiables, porque esto ayudó a mejorar el comportamiento del modelo, si este tipo de análisis se liga con la aplicación de técnicas estadísticas más apropiadas como la de los modelos de Box-Jenkins ARIMA y Principal Components Analysis (PCA). En realidad, el propósito principal de este estudio es para calcular una serie de análisis de regresión estadística, mismos que serán complementados con la aplicación de modelos de ARIMA y PCA para estratificar los valores de las concentraciones de ozono (valores altos y bajos). Consecuentemente, el estudio se dividió en dos etapas. Por ejemplo, la primera etapa describe la base de datos a través del uso de mediciones estadísticas básicas, relacionadas con la aplicación de análisis de regresión. La segunda etapa está sujeta al modelado de ARIMA y PCA, cuya tarea se hará en un futuro estudio.

Los datos de las variaciones espacio-temporales de 1-hora de ozono (O_3), óxidos de nitrógeno (NO y NO_2) y datos meteorológicos, como temperatura del punto de rocío, temperatura ambiental, humedad relativa, radiación solar, intensidad del viento, intensidad del viento resultante, dirección del viento resultante y ráfagas máximas (denotados por los términos Dew, T, Hum, Sun, WS, RWS, RWD, Gust, respectivamente) fueron obtenidos de una estación de muestreo ubicada en el Parque Chamizal, la cual está localizada en la línea divisoria entre El Paso, Texas (E. U. A.) y Ciudad Juárez, Chihuahua (México). Con esta información, este estudio obtuvo un análisis estadístico del ozono representativo para el área de Juárez y lugares circunvecinos, por medio de procesar las concentraciones de ozono a nivel del suelo de 1-hora. El modelado estadístico propuesto involucra los datos reales de ozono de 1-hora (variable de respuesta), los cuales fueron experimentados en regresión contra un grupo de covariantes químicos y meteorológicos, como NO, NO_2, temperatura de punto de rocío, temperatura ambiental, humedad relativa, intensidad del viento, intensidad del viento resultante, dirección del viento resultante y ráfagas máximas (tomadas como las variables independientes).

Básicamente, la metodología consistió en la selección del mejor modelo de regresión estadístico, es decir, por medio de ajustar varios modelos candidatos de regresión lineal múltiple y funciones de regresión polinomial. Finalmente, para complementar los procedimientos anteriores, el estudio incluyó los promedios aritméticos y las desviaciones estándar de los valores de ozono, así como también de NO, NO_2 y variables meteorológicas.

MATERIALES Y MÉTODOS

Los datos de esta investigación fueron obtenidos de la estación de muestreo del Parque Chamizal (EPA sitio numero 48-141-0044) localizada en la línea divisoria, entre Ciudad Juárez y , la cual está situada a una elevación de 1122.1 metros arriba del nivel del mar. Esta estación de muestreo es operada y mantenida por la de la oficina regional de Texas. El banco de datos usados para la estructuración del modelo de regresión de este estudio consistió de cerca de 47,000 casos correspondientes al período 1999-2003.

Los datos fueron procesados bajo la plataforma de la ventana del programa Excel (Visual Basic for Applications Language), y analizados, estadísticamente, usando el paquete de computadora Minitab.

El modelo de regresión básico usó el ozono (O_3), como la variable de respuesta o dependiente y una lista total de diez posibles variables independientes como el punto de rocío, NO, NO_2, temperatura ambiental, humedad relativa, radiación solar, intensidad el viento, intensidad del viento resultante y ráfagas máximas (NO, NO_2, Dew, T, Hum, Sun, WS, RWS, RWD, y Gust).

La metodología usada en este estudio estadístico del ozono troposferico, consistió de lo siguiente: Primeramente, el estudio ajustó un modelo de regresión lineal múltiple, por medio de incluir todas las diez variables independientes. Esto se hizo basándose la ecuación paramétrica de regresión lineal múltiple de la forma de:

$$Y_j = \beta_o + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} \ldots\ldots + \beta_k X_{kj} + \varepsilon_j \qquad (4)$$

Donde Y_j es el valor de la variable dependiente poblacional; β_o es el valor del intercepto en la ordenada; β_2, …, β_k son los valores de los coeficientes parciales de regresión; X_{1j}, X_{2j}, ……, X_{kj} son los valores de las variables paramétricas; y, ε_j es el factor de error asociado con las fluctuaciones de las variables aleatorias.

La función estadística de regresión lineal múltiple que emula a la función (4) es:

$$Y_j = b_o + b_1 X_{ij} + b_2 X_{2j} + b_3 X_{3j} + \ldots\ldots + b_k X_{kj} + e_j \qquad (4a)$$

Donde Y_j es el valor de la concentración del ozono; b_o es el intercepto; b_1, b_2, b_3, etc., son las pendientes parciales del plano asociado;
X_1 es la temperatura del punto de rocío; X_2 es el NO; X_3 es el NO_2; X_4 es la temperatura ambiental y, así sucesivamente. Enseguida, el procedimiento usado en la selección del modelo exploró una ecuación de regresión cuadrática polinomial, la cual incluyó todas las diez variables. Esto se hizo basándose en la ecuación paramétrica abreviada de abajo:

$$Y_j = b_o + \sum_{i=1}^{k} \beta_i X_{ij} + \sum_{i=1}^{k} \gamma_i X_{ij}^2 + \varepsilon_j \qquad (5)$$

La función estadística de regresión lineal múltiple que emula a la función (5) es:

$$Y_i = b_o + b_1 X_{ij} + b_2 X^2_{ij} + \ldots. + b_j X_k + b_k X^2_k + e_j \qquad (5^a)$$

Donde *Yi* es la concentración de ozono troposférico, *bo* es el intercepto, b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9 y b_{10}, son los gradientes parciales del plano asociado X_1, etc.

El siguiente paso en el proceso de la estructuración y refinación del modelo consistió en eliminar los datos asociados con los valores residuales extremos (valores atípicos), para poder obtener un mejor ajustamiento de los valores reales de las concentraciones de ozono. Esto se hizo usando una serie de modelos de regresión lineales múltiples y modelos de regresión polinomiales

cuadráticos que excluyeron los datos asociados con los valores atípicos. Para este fin, el procedimiento eliminó todos los residuos estandarizados con valores absolutos ≥ 3, es decir, siguiendo el procedimiento señalado por Neter et al. (1996), el cual elimina los valores atípicos extremos. Siguiendo este razonamiento, el estudio ajustó un modelo lineal múltiple y un modelo polinomial cuadrático, los cuales excluyeron los datos asociados con los referidos valores atípicos.

El diagnóstico usado en la selección del modelo estadístico óptimo se hizo a través de un profundo análisis objetivo y subjetivo, es decir, con el uso de varios criterios estadísticos. Estos criterios incluyeron el coeficiente de determinación múltiple R^2, s, PRESS (del inglés, prediction error sum of squares o suma de cuadrados de error de predicción) y factor de inflación de varianza (del inglés, VIF), para revisar por problemas de multicolinealidad. Posteriormente, para asegurarse de que no hubiera problemas de extrapolación, es decir, usando los valores de las variables de predicción fuera de la cáscara regresiva, los elementos diagonales del matriz de sombrero fueron calculados para detectar una posible extrapolación escondida Sin embargo, no se obtuvieron valores de extrapolación escondidas en ningún caso. Finalmente, se hicieron pruebas de normalidad e histogramas de la distribución de residuales para complementar el procedimiento de evaluación de los modelos candidatos.

Una descripción detallada de los diagnósticos estadísticos usados para la selección del modelo de regresión superior es como sigue:

1. Coeficiente de determinación múltiple (R^2) y coeficiente de determinación múltiple ajustado $R^2_{(ajustado)}$, respectivamente, donde R^2 representa el porcentaje de variación de la variable dependiente (concentraciones de ozono a nivel del suelo) explicado por el juego de variables independientes (NO, NO_2 y de todos las demás covariantes meteorológicos). Este criterio R^2 mide la fuerza de la relación lineal entre los componentes del modelo (concentraciones de ozono y sus covariantes). Entre más alto sea el valor de R^2 las concentraciones de ozono observadas estarán mejor ajustadas por el modelo de regresión aplicado. No obstante, es necesario aclarar que un valor alto de R^2, no necesariamente indicaría un buen modelo de regresión de ozono, como tampoco lo indicaría un valor de R^2 pequeño. Por otro lado, el coeficiente $R^2_{(ajustado)}$, es una versión ajustada de R^2 que busca remover las sobreestimaciones debidas a un tamaño de muestra pequeño.
2. Error estándar de lo estimado (s). El valor de s es un diagnóstico estadístico importante el cual involucra las diferencias entre los valores reales del modelo y los valores del ozono pronosticados. Entre más pequeño sea el valor de s, mejor será el modelo de ozono seleccionado. Sin embargo, entre más grande sea el valor de s, más dispersión habrá en los puntos, con respecto al plano de regresión.
3. PRESS (del inglés, prediction error sum of squares o suma de cuadrados del error de predicción). Este criterio, que también se le suele llamar residuales eliminados, se usó para validar el modelo de regresión de ozono en términos de validación cruzada de errores o en términos de predicción. En este instante, es de notarse que entre más pequeño sea el valor de PRESS mejor será el modelo seleccionado.
4. Criterio Cp de Mallows. Este criterio está relacionado con el promedio total del cuadrado del error de los valores *n* ajustados para cada subconjunto de modelos de regresión. El modelo óptimo seleccionado tiene un valor de Cp cercano a (p + 1), donde, p es el número de variables regresoras. Un valor Cp mayor que (p + 1) indica que el modelo tiene variables innecesarias, mientras que un Cp menor que (p + 1) indica que hubo omisión de variables importantes. En este estudio, esta estadística se usó para determinar el número óptimo de variables independientes incluidas en el modelo de ozono más apropiado.
5. Factor de inflación de varianza (VIF). (VIF del inglés, variance inflation factors). Este factor se usa para revisar los problemas de multicolinealidad, es decir, cuando las variables regresoras están altamente correlacionadas entre si (coeficientes de regresión inflados). En este renglón, Neter et al. (1996) afirman que un promedio máximo de valores de VIF en exceso de 1.0 significa que la colinealidad pueda estar influenciando las estimaciones de los cuadrados mínimos. Sin embargo, estos autores afirman que, el hecho de que algunos regresores estén linealmente asociados, semejante condición no necesariamente inhibe un buen ajuste, ni tampoco afecta las inferencias acerca de las predicciones del promedio, siempre y cuando estas inferencias estén dentro de la región de las observaciones originales (lo cual se observó en este estudio). Otros diagnósticos informales mencionados por los mismos autores, indican que pueden ocurrir grandes cambios en los coeficientes de regresión estimados, cuando un regresor es agregado o eliminado o cuando una observación es alterada o eliminada. Similarmente, estos

autores mencionan otros diagnósticos relacionados con la multicolinealidad, es decir, cuando los coeficientes de regresión tienen signos algebraicos opuestos a lo que se esperaría de experiencia o lógica a posteriori (situación que no ocurrió con el modelo superior obtenido en este estudio). Por otra parte, Montgomery et al. (2001) afirma que si los factores VIF son mayores que 10, esto implica problemas graves de multicolinealidad. Además, estos investigadores afirman que si los coeficientes de regresión tienen signos incorrectos, esto pueda deberse a que el intervalo de alguno de los regresores es demasiado pequeño. Esta condición también sugiere que, la aparición de signos incorrectos, pueda deberse a que no se han incluido variables regresoras importantes en el modelo. Esta condición se puede deber a que hay multicolinealidad o, bien, que se han cometido errores de cómputo. Análogamente, Pfaffenberger et al. (1987), discuten el problema de la multicolinealidad basándose en los análisis de VIFs. Estos investigadores argumentan que, si el valor máximo del VIF es mayor que 1, entonces, hay problemas graves de colinealidad (situación que no ocurrió en la selección del modelo superior de este estudio).

6. Elementos diagonales de matriz de sombrero (h_{ii}). Para revisar por extrapolación oculta, es decir, que los cálculos del modelo superior estuvieran dentro de la cáscara de variables regresoras, los elementos diagonales de la matriz de sombrero [$H = X'(XX)'X^{-1}$] (Cook, 1979) se obtuvieron usando el paquete de computadoras Minitab software. En ésta instancia, un h_{ii} para cada caso se calculó y se comparó contra el h_{max} obtenido para cada modelo aplicado.

En resumen, la metodología aplicada en este estudio estadístico consistió en probar, primeramente, la inclusión de todas la variables independientes, por medio de probar una serie de modelos de regresión lineal múltiple y de regresión polinomial, es decir, incluyendo todas la variables independientes (ver ecuaciones 5 y 6, respectivamente). De esta manera, se probó un modelo de regresión cuadrático que incluyó todas las variables independientes. Enseguida, el estudio examinó los valores de los criterios estadísticos R^2, $R^2_{(ajustado)}$, s, PRESS y de Cp, para determinar el mejor modelo de regresión. Siguiendo este procedimiento, la base de datos se escudriñó por medio de eliminar todos los casos que mostraran valores extremos o atípicos que pudieran afectar, adversamente, el ajustamiento de las concentraciones de ozono observadas. Posteriormente, el estudio ajustó modelos de regresión lineal múltiple y modelos cuadráticos polinomiales usando los algoritmos de regresión Stepwise y Best-ubset, es decir, en conjunción con un juicio estadístico objetivo. Esto se hizo para eliminar o conservar las variables basando el criterio en los resultados de los diagnósticos objetivos estadísticos. Además, este procedimiento fue apoyado por medio de examinar los factores de inflación de varianza (VIF) de cada uno de los modelos probados para seleccionar el modelo superior. Finalmente, los h_{ii} de cada caso de los modelos probados se calcularon y se analizaron acordemente.

Para complementar los procedimientos anteriores, este estudio estimó el promedio anual (μ), la desviación estándar (σ) y el coeficiente de variación [$CV = (\sigma/\mu)100$], para el ozono, y para los datos químicos y meteorológicos correspondientes al periodo 1999-2003 de este estudio.

RESULTADOS Y DISCUSIÓN

En la selección crítica del mejor modelo estadístico, el estudio se inició por medio de probar un modelo de regresión lineal múltiple que incluyó todas las variables y todos los valores observados. Por ejemplo, el Modelo 1 de la **tabla 1** muestra los resultados de esta aplicación. Enseguida, el proceso de selección del modelo óptimo exploró una función de regresión cuadrática (Modelo 2), que incluyó las mismas variables y casos incluidos en el Modelo 1. Los resultados del Modelo 2 se muestran en la **tabla 2**.

El siguiente paso consistió en ajustar un modelo de regresión lineal múltiple, el cual excluyó los valores residuales extremos o atípicos. Prosiguiendo de esta manera, y después de probar varios modelos de regresión, por medio de revisar los valores de los diagnósticos estadísticos, se obtuvo el Modelo 3. Este Modelo 3 incluyó todas las variables regresoras, excepto la intensidad del viento, humedad relativa y ráfagas máximas. Los resultados de este Modelo se muestran en la sección media de la **tabla 1** (ver Modelo 3).

TABLA 1. Resultados de los modelos de regresión lineal múltiple.

Regresor	Modelo 1			Modelo 3			Modelo 5		
	B_I	(Se)	VIF	B_I	$(Se)_I$	VIF	B_I	(Se)	VIF
Desconocido	3.7E+00	4.9E-01	N/A	7.9E+00	3.1E-01		5.7E+00	3.0E-01	N/A
Dew	-1.2E-01	7.9E-03	4.5	-1.4E-01	4.6E-03	1.6	-1.3E-01	4.4E-03	1.6
NO	-3.2E-02	1.3E-03	1.6	-2.6E-02	1.2E-03	1.6	-2.5E-02	1.2E-03	1.6
NO_2	-4.9E-01	6.0E-03	1.9	5.8E-01	5.8E-03	1.8	-5.3E-01	5.7E-03	1.9
T	3.9E-01	6.9E-03	4.9	4.2E-01	3.9E-03	1.7	3.9E-01	3.9E-03	1.8
Sun	1.3E+01	1.4E-01	1.3	1.4E+01	1.3E-01	1.2	1.3E+01	1.3E-01	1.3
RWS	-2.9E+00	8.4E-02	57.5	3.7E-01	1.3E-02	1.5	-7.1E-01	2.5E-02	5.7
RWD	1.8E-03	6.0E-04	1.3	3.2E-03	6.0E-04	1.3	1.9E-03	6.0E-04	1.3
Gust	4.7E-01	1.9E-02	8.8	N/A	N/A	N/A	7.4E-01	1.5E-02	6.1
WS	2.7E+00	1.0E-01	78.9	N/A	N/A	N/A	N/A	N/A	N/A
Hum	2.0E-04*	5.4E-03	3.9	N/A	N/A	N/A	N/A	N/A	N/A

* Estadísticamente insignificante

Siguiendo el procedimiento anterior, se calculó una ecuación de regresión cuadrática para ajustar el modelo 4. Este modelo de regresión excluyó los valores residuales extremos o atípicos. Este modelo también excluyó la intensidad del viento, la humedad relativa y las ráfagas máximas. Los resultados del modelo 4 se muestran en la sección derecha de la **tabla 2**. En este caso, los valores del coeficiente de determinación R^2 y de las estadísticas s y PRESS sugieren que el mejor modelo es el Modelo 4. Sin embargo, examinando la **tabla 2**, se observa que los valores del factor de inflación de varianza (VIF) de este modelo son sustancialmente altos, cuyos rangos son de 2.5 a 167.8, es decir, con un promedio de VIF de 40.0. De acuerdo a Meter et al. (1996), un valor de VIF en exceso de 10 indica que la multicolinialidad pudiera estar afectando adversamente el coeficiente de la ecuación de regresión asociado a la variable dependiente. Estos investigadores también contienden que un valor promedio del VIF considerablemente más grande que 1.0 indica problemas de colinealidad severos (el cual es el caso en el ajustamiento del Modelo 4). A más de, al estudiar los signos algebraicos de los coeficientes de regresión de este Modelo 4, el signo negativo enfrente de la variable regresora, es decir, de la temperatura, no es correcto (ver **tabla 2**, Modelo 4). Esto se debe a que, la temperatura debería estar contribuyendo una cantidad positiva a la producción de ozono a nivel suelo, de acuerdo a un razonamiento a posteriori.

TABLA 2. Resultados de los modelos de regresión cuadráticos.

Regresor	Modelo 2			Modelo 4		
	B_I	$(Se)_I$	VIF	B_I	$(Se)_I$	VIF
Desconocido	2.5E+01	8.6E-01	N/A	3.5E+01	8.6E-01	N/A
Dew	-2.6E-01	1.7E-02	22.9	-2.2E-01	1.7E-02	24.0
NO	-1.3E-01	2.7E-03	8.0	-1.3E-01	2.6E-03	7.9
NO_2	-8.7E-01	1.4E-02	10.8	-9.2E-01	1.4E-02	11.8
T	-1.9E-01	2.4E-02	64.1	-2.2E-01	2.4E-02	69.5
Sun	1.4E+01	4.1E-01	13.4	1.5E+01	3.9E-01	13.3
WS	5.7E+00	1.9E-01	310.2	N/A	N/A	N/A
RWS	-5.9E+00	1.4E-01	181.8	2.9E-01	3.3E-02	10.4
RWD	1.6E-03*	2.8E-03	29.8	-5.1E-03	2.7E-03	29.9
Gust	9.3E-01	4.7E-02	60.9	N/A	N/A	N/A
$(Dew)^2$	2.0E-03	2.0E-04	22.6	1.7E-03	2.0E-04	22.6
$(NO)^2$	3.0E-04	1.0E-05	5.9	3.0E-04	1.0E-05	5.8
$(NO_2)^2$	8.1E-03	2.0E-04	9.3	8.0E-03	2.0E-04	10.5
$(T)^2$	4.1E-03	2.0E-04	63.7	4.2E-03	2.0E-04	64.6
$(Sun)^2$	-1.6E+00	3.6E-01	13.4	-1.8E+00	3.4E-01	13.4
$(WS)^2$	-2.2E-01	1.0E-02	430.0	1.1E-01	6.0E-03	167.8
$(RWS)^2$	2.3E-01	8.9E-03	302.9	-1.4E-01	6.0E-03	150.1
$(RWD)^2$	-1.0E-05*	1.0E-05	30.4	2.0E-05	1.0E-05	30.5
$(Gust)^2$	-1.3E-02	1.0E-03	49.1	6.6E-03	4.0E-04	7.5
$(Hum)^2$	N/A	N/A	N/A	-2.0E-04	4.0E-05	2.5

* Estadísticamente insignificante.

Nota: NO = Monóxido de nitrógeno, NO_2 = Dióxido de nitrógeno, Dew = Temperatura de punto de rocío, T = Temperatura ambiental, Sun = Radiación solar, WS = Intensidad del viento, RWS = Intensidad del viento resultante, RWD = Dirección del viento resultante, Gust = Ráfagas máximas, Hum = Humedad relativa

Finalmente, el estudio ajustó un modelo de regresión lineal múltiple (Modelo 5). Este Modelo 5 excluyó los valores residuales extremos, la intensidad del viento y la humedad relativa. Los resultados de este Modelo 5 se muestran en la **tabla 1**. A excepción del Modelo 5, los valores inflados de la varianza (VIF) de los otros modelos mostrados en las **tablas 1** y **2** indican resultados similares como aquéllos asociados al Modelo 4. En este caso, sin embargo, el VIF más grande del Modelo 5 es igual a 6.1, el cual es menor que 10.0. Además, el promedio de estos valores de VIF (2.7) no es muy alto comparado con el valor de 1. Encima de, los signos algebraicos de los coeficientes de la ecuación de regresión del Modelo 5 están de acuerdo a lo que se esperaría de una lógica *a posteriori*. Más adelante, los elementos diagonales asociados a la matriz de sombrero (h_{ii}) de este análisis, mostraron que no hubo extrapolaciones calculadas en el Modelo 5. Adicionalmente, las **figuras 1** y **2** mostraron que los valores residuales del Modelo 5 siguieron a una distribución normal, lo cual es requerido por la teoría básica asociada con el modelado de regresión. Todos estos análisis sugirieron que el Modelo 5 es el mejor modelo estadístico. Otra razón por la cual esta investigación concluyó que el Modelo 5 es el más apropiado, para ajustar los datos del ozono, se debe a qué, este Modelo es simple y tiene el número más apropiado de variables regresoras, al juzgar por el razonamiento estadístico aplicado en esta investigación. En este respecto, se observa que modelos de regresión con muchas variables independientes, no necesariamente los hace mejor, porque esto no ayuda a economizar tiempo, dinero y recursos.

TABLA 3. Criterios estadísticos para los mejores cinco modelos de regresión obtenidos.

Modelo ajustado	R^2 (%)	$R^2_{(ajustada)}$ (%)	s	PRESS (10^6)	Cp	VIF Promedio	VIF Máximo
Modelo 1[a]	67.7	67.7	10.8	5.1	10.0	16.5	78.9
Modelo 2[b]	71.1	71.1	10.2	4.6	20.0	90.5	430.0
Modelo 3[c]	68.2	68.2	10.4	4.7	8.0	1.5	1.7
Modelo 4[d]	71.8	71.7	9.8	4.2	18.0	37.8	167.8
Modelo [e]	70.0	69.9	10.1	4.4	9.0	2.6	6.1

[a] Regresión lineal usando todos los casos de los valores originales.
[b] Regresión cuadrática usando todos los casos de los valores originales
[c] Regresión lineal usando todas las variables regresoras, excepto WS, Hum, y Gust. Los casos atípicos fueron eliminados
[d] Regresión cuadrática usando todas las variables regresoras, excepto WS, Hum, y Gust. Los casos atípicos fueron eliminados
[e] Regresión lineal usando todas las variables regresoras, excepto WS, Hum, y Gust. Los casos atípicos fueron eliminados.

La **tabla 3** muestra un resumen de los criterios estadísticos que fueron calculados para seleccionar el mejor modelo de regresión, entre todos los modelos que se probaron. De los criterios dados en la **tabla 3**, a pesar de que los valores de los elementos diagonales de la matriz de sombrero (h_{ii}) no se mostraron explícitamente, no hubo problemas de extrapolación en ninguno de los modelos de regresión investigados.

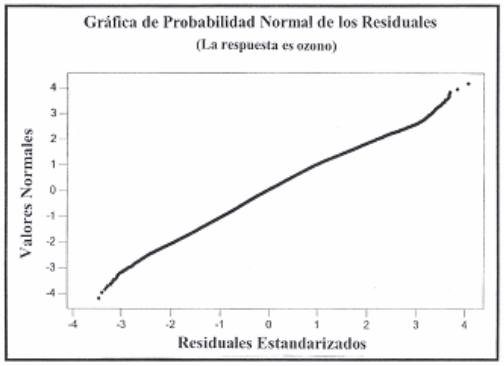

Figura 1. Prueba de normalidad para la distribución de los valores residuales del Modelo 5

Para complementar los procedimientos anteriores, este estudio estimó el promedio anual (μ), la desviación estándar (σ) y el coeficiente de variación [$CV = (\sigma/\mu)100$], para el ozono, y para los datos químicos y meteorológicos correspondientes al periodo 1999-2003 de este estudio. La **tabla 4** muestra que, a excepción de las variables independientes Sun y Hum, todas las demás variables exhibieron valores CV relativamente bajos. Derivado de estas premisas, se pueden derivar dos situaciones: (1) debido a los valores bajos del coeficiente de variación (CV) se concluye que los datos coleccionados para el periodo de estudio, son relativamente, homogéneas. lo cual apoya, aun más, la validez del mejor modelo obtenido en este estudio (Modelo 5); y (2) las variables Sun y Hum no son tan homogéneas para estos juegos de datos. Tal vez esta sea la razón por la cual se obtuvo el Modelo 5 sin la contribución de la variable regresora humedad relativa.

TABLA 4. Promedio aritméticos anuales para la variable de respuesta (ozono) y para las variables regresoras.

Año	Variable de respuesta	Variables regresoras								
	O_3	Dew	NO	NO_2	T	Sun	RWS	RWD	Gust	WS	Hum
1999	26.2	34.4	24.6	22.5	68.4	0.3	7.2	186.1	15.4	7.9	42.0
2000	25.8	35.1	22.0	23.6	68.3	0.2	6.8	178.4	14.8	7.5	38.3
2001	25.5	34.8	21.4	21.0	67.1	0.3	7.0	178.7	14.8	7.6	35.4
2002	27.3	33.7	21.5	21.4	67.1	0.3	6.8	171.0	14.4	7.4	35.4
2003	26.5	33.3	20.5	20.2	67.7	0.3	6.9	171.9	14.7	7.5	32.9
Estadísticas anuales											
μ	26.3	34.3	22.0	21.7	67.7	0.3	6.9	177.2	14.9	7.6	36.8
σ	0.6	0.7	1.4	1.2	0.6	0.0	0.1	5.5	0.3	0.2	3.1
$CV = (\sigma/\mu)100$	2.4	2.0	6.3	5.5	0.8	12.1	2.1	3.1	2.2	2.3	8.4

Finalmente, al examinar **la tabla 3**, se puede afirmar que, el ajustamiento del Modelo 5 (modelo de regresión aplicado, sin la presencia de observaciones inusuales para todas las variables originales excepto WS and Hum), aparece como el mejor modelo de regresión. Esto es apoyado al juzgar por los criterios estadísticos (R^2, $R^2_{(ajustada)}$, s, PRESS, Cp y VIF) y demás diagnósticos estadísticos usados en este estudio. Un refinamiento posterior de estos procedimientos, con la inclusión de la metodología GAM se hará en otro futuro estudio.

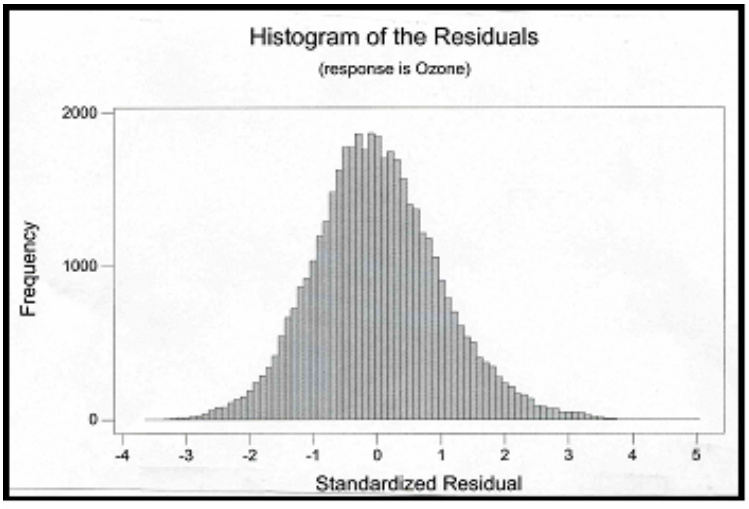

Figura 2. Histograma de los valores residuales del Modelo 5.

REFERENCIAS

Brown T. L., H. E. LeMay, Jr. y B. E. Busten. (2000). *Chemistry: The Central Science*. Octava edición. Prentice Hall. Upper Saddle River, New Jersey 07458, 1000 p.

Cook, R. D. (1979). Influential Observations in Linear Regression. J. Am. Stat. Assoc., 74, pp. 169-174.

Davis J. M. y P. Speckman. (1998). A Model for Predicting Maximum and 8-Hour Average Ozone in Houston. Department of Marine, Earth and Atmospheric Sciences and Plant Pathology, North Carolina State University, Raleigh, NC, USA.

Eder B. K.; J. M. Davis y P. Bloomfield. (1993). A Characterization of the Spatio-temporal Variability of Non-Urban Concentrations Over the Eastern United States. Atmospheric Environment, Vol. 27A, No. 16, pp. 2645-2668.

Hubbard M. C. y W. G. Cobourn. (1998). Development of a Regression Model to Forecast Ground-Level Ozone Concentrations in Louisville, KY. Department of Mechanical Engineering, Speed Scientific School, University of Louisville, Kentucky, U. S. A.

Libiseller C. y A. Grimvall. (2003). Model Selection for Local and Regional Meteolological Normalisation of Background Concentrations of Tropospheric Ozone. Department of Mathematics, Linköping University, SE-58183 Linköping, Sweden.

Neter J.; M. H. Kutner, C.J. Nachtsheim y W. Wasserman. (1996). *Applied Linear Regression Models*. Third ed. McGraw-Hill Companies, Inc. 720 p.

Pfaffenberger R. C. Y J. H. Patterson (1987). *Statistical Methods*. Third ed. Richard D. Irwin, Inc. Homewood, Illinois, 1246 p.