Análisis de series de tiempo univariante aplicando metodología de Box-Jenkins para la predicción de ozono en la ciudad de Cali, Colombia

Mauricio Jaramillo Ayerbe[a, *], *Daniel Enrique González Gómez*[a], *María Eugenia Núñez Cabrera*[a], *Gloria Esperanza Portilla*[a], *Jesús Heriberto Lucio García*[b]

[a]Grupo de Producción más Limpia. Facultad de Ingeniería. Pontificia Universidad Javeriana, Sede Cali. Cali, Colombia.

[b]Departamento de Física, Universidad de Burgos, Burgos, España.

(Recibido el 21 de octubre de 2005. Aceptado el 28 de agosto de 2006)

Resumen

En este trabajo se presentan los resultados de la modelación para la predicción a corto plazo de la concentración de ozono troposférico en la zona urbana de la ciudad de Cali, Colombia, mediante el análisis univariante de series de tiempo. El método se aplicó a una serie de 2.496 datos horarios de concentración de ozono, obtenida en una de las estaciones de la Red de Monitoreo de Calidad del Aire (RMCA) de Cali, correspondiente a 104 días consecutivos durante el período abril a julio de 2003. Los datos de los primeros 93 días se utilizaron para la estimación del modelo, y los de los 11 días restantes para la validación del mismo. El modelo propuesto podría ser utilizado por las autoridades ambientales de la región para predecir y alertar a la población sobre posibles episodios de altas concentraciones de ozono que puedan poner en riesgo la salud pública con hasta 8 horas de anticipación.

---------- *Palabras clave:* contaminación atmosférica, calidad del aire, ozono urbano, análisis de series de tiempo, análisis univariado.

Univariant time series analysis applied to ozone forecasting in the city of Cali, Colombia

The modelling results of tropospheric ozone concentration in the urban area of Cali, Colombia, suitable for short term forecasting, are presented. Results were obtained by using an univariate time series analysis. The method was applied to

* Autor de correspondencia. Teléfono: +57-2-321 83 52, fax: +57-2-555 28 23. correo electrónico: mjaramil@puj.edu.co. (M. Jaramillo)

a series of 2496 hourly ozone concentration data from one of the city's air quality monitoring network stations. Data were collected from April to July of 2003. A total of 104 consecutive days were covered: the first 93 days were used for model estimation, and the remaining 11 days for model validation. This technique can be used to predict (up to 8 hours) in advance high-ozone levels, allowing the environmental authorities to issue alerts to the population for possible air-quality impact on health.

---------- *Key words:* atmospheric pollution, air quality urban ozone, time series analysis, univariate time series analysis.

Introducción

Las predicciones de la concentración de ozono mediante métodos cualitativos y cuantitativos se han convertido en una importante tarea a llevar a cabo por las entidades y organismos reguladores, siendo cada vez más común enfrentar la necesidad de tomar una decisión o evaluar el impacto que tendrá un contaminante en la calidad del aire y en la salud.

En Cali, una ciudad de más de dos millones de habitantes en el suroccidente de Colombia, la autoridad ambiental local (Departamento Administrativo de Gestión del Medio Ambiente, DAGMA) opera las siete estaciones de la Red de Monitoreo de Calidad del Aire (RMCA) donde se mide la concentración de algunos contaminantes y también parámetros meteorológicos, pero no se realizan sondeos de perfiles atmosféricos. El aire es una preocupación ambiental creciente debido a la presencia de concentraciones considerables de contaminantes fotoquímicos como ozono (O_3) en la zona urbana, resultado del crecimiento mismo de la ciudad, sus características ambientales, la industrialización y la circulación de vehículos que contribuyen al incremento de las emisiones y al deterioro de la calidad del aire.

La obligación social de informar a la población sobre la posibilidad de superar límites permisibles de contaminantes y también de realizar predicciones con fines preventivos, hacen del análisis estadístico una valiosa herramienta para fortalecer el grado de certeza de las proyecciones.

En el presente estudio se aplicó el análisis univariado de series de tiempo a las mediciones de la red de monitoreo para identificar un modelo que permita predecir niveles de ozono a corto plazo y realizar con facilidad predicciones a partir de un número reducido de datos.

Metodología

La técnica estadística de análisis univariado de series de tiempo se ha utilizado en diferentes áreas del conocimiento como la economía, la física, la medicina entre otras y recientemente en el análisis de datos de calidad del aire [1, 2]. En el presente trabajo seguimos lo que se conoce como el enfoque Box-Jenkins para el análisis de los datos y la identificación de un modelo apropiado. Las siguientes son las etapas generales consideradas en el estudio:

Selección de datos

Los datos utilizados para la modelación se obtuvieron de los suministrados por el DAGMA y corresponden a una serie de datos continuos de 104 días tomados a partir del 1 de abril de 2003 a las 11:00 horas hasta el 14 de julio a las 10:00 horas locales. La información recibida del DAGMA incluye datos de emisiones de SO_2, NO, NO_2, CO, O_3, PM_{10} y radiación solar, temperatura, humedad, dirección del viento, velocidad del viento para siete estaciones de monitoreo, así como el registro de novedades debido a fallas eléctricas, calibración o mantenimiento de los equipos. De estas estaciones se seleccionó la que se encuentra en el Centro de Diagnóstico Automotor (CDA), localizada al norte de la ciudad, por presentar datos completos de las variables a analizar en el estudio.

Del conjunto de datos disponibles para esta estación se seleccionaron 2.496 datos horarios consecutivos correspondientes a concentraciones atmosféricas de ozono (O_3), expresadas en partes por mil millones por volumen (p. p. b.). La serie de datos de ozono, correspondiente a 104 días, fue dividida en dos partes: la primera incluye la información de 93 días (2.232 datos horarios) y se utilizó para la estimación del modelo y la segunda, con una duración de 11 días (264 datos horarios), destinada a la validación del mismo. Estos datos presentaron variabilidad considerable en algunos casos, lo que hizo necesaria la aplicación de técnicas de ajuste de datos con el fin de controlar su varianza.

Análisis de datos

Para la depuración de datos se realizó un análisis global para identificar valores extremos o extraños examinando sus posibles causas por falla en

el método de análisis o de registro y realizando correcciones a estos casos. Este procedimiento incluyó la graficación de los datos por horas, días y semanas para identificar puntos extremos o ausencia de valores (valores atípicos), para visualizar la tendencia de los mismos y eliminar aquellos valores que difieren del límite de detección de los dispositivos de monitoreo.

Se confrontó con la base de datos de registro de novedades generada por el DAGMA para explicar la presencia de estos valores atípicos, y para aquellos valores donde no fue posible encontrar justificación, se utilizó interpolación lineal clásica entre los datos adyacentes. En algunos casos en que faltaban varios datos horarios consecutivos, éstos se reemplazaron por el promedio total de los datos correspondientes a cada hora del día específico, previa verificación de que existiese correlación. Finalmente se realizó este procedimiento de ajuste a un total de 18 datos.

Estabilización de varianza y desestacionalización

Con el fin de estabilizar la serie en varianza y facilitar su modelación se aplicó en una primera instancia la transformación Box-Cox [3] con parámetros $\lambda = -0{,}09$ y $c = 0$ y luego, a la serie resultante, un proceso de desestacionalización utilizando transformada rápida de Fourier con período de 24 horas.

Selección preliminar del modelo

Una vez estabilizada la serie se procedió a estudiar la presencia de regularidades para identificar posibles modelos matemáticos. Para ello se calculó la función de autocorrelación tanto simple como parcial y se compararon con los diferentes patrones que suministra la metodología Box-Jenkins, que son típicos de las diferentes funciones generadoras de datos, seleccionando los modelos que mejor se ajusten o adecuen a la forma de las funciones de autocorrelación obtenidas.

Después de la selección de los mejores modelos, se estimaron los coeficientes de los mismos y se procedió a efectuar un análisis de los residuos (diferencia entre el valor realmente observado y el valor previsto por el modelo), verificando la presencia de un comportamiento estacionario o de ruido blanco para seleccionar el modelo que presente el ajuste más adecuado.

Resultados y discusión

El esquema general del modelo es el siguiente:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} \quad (1)$$

El valor de la variable de interés (concentración de ozono) y_t en el instante t se expresa en función de sus valores previos y de valores aleatorios ε presentes y anteriores con coeficientes ϕ_i y θ_i que se determinan numéricamente ajustando el modelo a la serie. Esta es la forma de los modelos ARMA(p,q), constituidos por una combinación de p términos AR (proceso autorregresivo) y términos q términos MA (proceso de medias móviles) [4]. p y q son los valores máximos de *retardo* en los términos autorregresivos y de media móvil, respectivamente.

Una vez definido el modelo y estimados sus coeficientes, se restablecieron las características originales de la serie de datos, que fue transformada para inducir estacionalidad.

La estimación del modelo fue realizada con la ayuda del *software EViews5* y del lenguaje *Matlab 7*.

Análisis gráfico y exploratorio

Las concentraciones de ozono troposférico están determinadas principalmente por sus precursores, como óxidos de nitrógeno y compuestos orgánicos volátiles, y por la radiación solar [5] La serie completa, que se muestra en la figura 1, corresponde a 2.496 datos horarios (104 días), con concentración

media de 19 p. p. b., un mínimo de 1,9 y un máximo de 81. Este comportamiento es estable a través de las horas en los días analizados. El ozono presenta valores máximos alrededor de las 14 a 15 horas, que siguen a las horas de mayor intensidad de radiación solar, y menores concentraciones en las horas de la madrugada debido a la ausencia de radiación, como se puede apreciar en la figura 2, en donde el eje vertical muestra concentración horaria media de ozono en p. p. b. (partes por mil millones por volumen) y el eje horizontal la hora local, para los primeros 100 días de la serie.

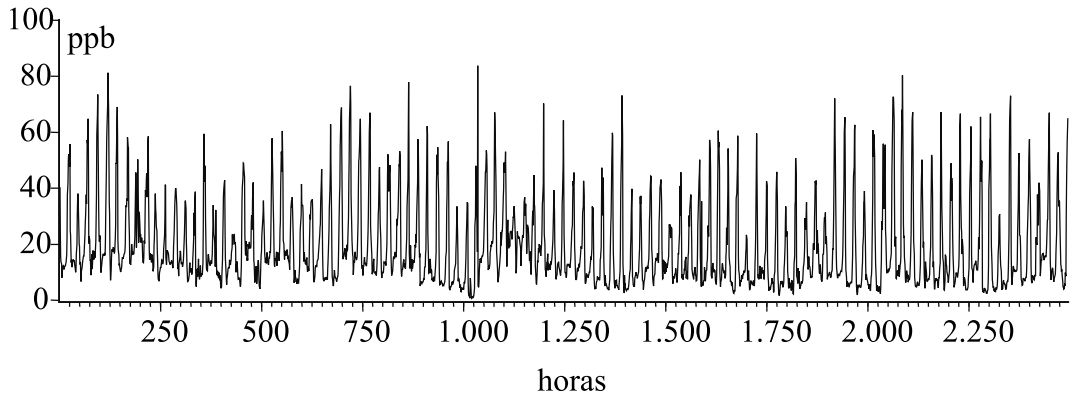

Figura 1 Serie completa correspondiente a 2.496 datos (104 días) de ozono obtenidos en estación Centro de Diagnóstico Automotor

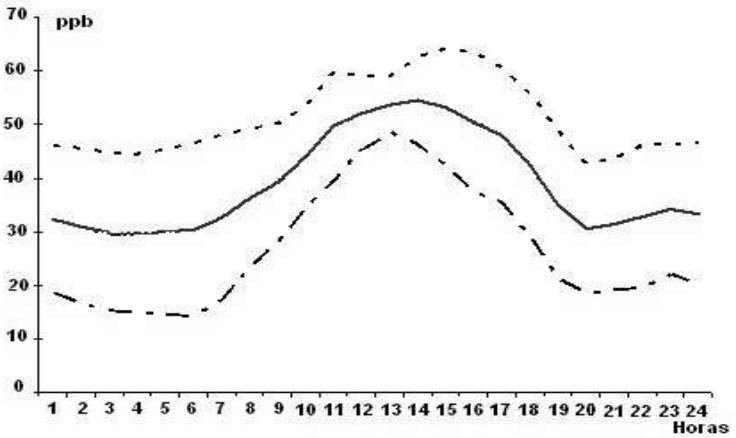

Figura 2 Variación diurna media de la concentración de ozono para los 100 días iniciales en p. p. b. versus hora local. El intervalo de confianza corresponde al 95%

Antes de establecer el patrón de generación de datos correspondiente a la serie, se realizan pruebas para detectar la presencia de raíces unitarias [4]. Esta detección es importante por cuanto es preciso comprobar el comportamiento estacionario de la serie, condición necesaria para la aplicabilidad de la metodología Box-Jenkins.

Con este fin, se sigue el proceso descrito por Enders [6] para la detección de raíces unitarias utilizando la prueba de Dickey Fuller Aumentada (ADF). Los resultados permiten concluir que la serie de datos de ozono es una serie estacionaria (valor p = 0,000), es decir, no requiere de ser diferenciada para su modelación mediante la metodología Box-Jenkins.

Identificación del modelo

Para la identificación del modelo se seleccionaron los primeros 2.232 datos horarios de la serie (93 días). Los 264 datos horarios de los 11 días restantes se reservaron para la validación y se utilizó la función de autocorrelación (ACF) y la función de autocorrelación parcial (PACF) de la serie transformada. Los valores para los cuales los coeficientes son significativamente diferentes de cero, indican el orden tanto autorregresivo (AR) como de media móvil (MA) del modelo, para lo cual se puede estimar el intervalo del 95% para estos coeficientes descritos en la ecuación 2:

$$-2/\sqrt{T} \le r(s) \le 2/\sqrt{T} = -0{,}04233 \le r(s) \le 0{,}04233$$

para número de datos T = 2.232 (2)

Es posible que un coeficiente que se encuentre por fuera de este intervalo sea diferente de cero al realizarle una prueba con hipótesis nula $H_O : \rho_s = 0$ [6].

Los valores que se encuentran fuera de este intervalo para la serie en estudio se muestran en la tabla 1.

Tabla 1 Coeficientes de autocorrelación (AC) y autocorrelación parcial (ACF) significativos

Retardo	*AC*	*ACP*
1	0,846	0,846
2	0,685	
3	0,566	0,055
4	0,467	
5	0,408	0,088
22	0,370	
23	0,385	
24	0,395	

Los resultados obtenidos utilizando el programa EViews5 sugieren que puede tratarse de un modelo ARMA(5,5) o ARMA(24,5).

Selección del modelo y estimación de los parámetros

De acuerdo con los valores de las correlaciones detectadas como significativas se evaluaron cuatro posibles modelos generadores de la serie. En todos los casos se trata de modelos ARMA(24,4).

Modelo 1 $$(o3d)_t = \phi_1(o3d)_{t-1} + \phi_2(o3d)_{t-2} + \phi_{24}(o3d)_{t-24} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_4\varepsilon_{t-4}$$

Modelo 2 $$(o3d)_t = \phi_1(o3d)_{t-1} + \phi_{24}(o3d)_{t-24} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_4\varepsilon_{t-4}$$

Modelo 3 $$(o3d)_t = \phi_1(o3d)_{t-1} + \phi_{22}(o3d)_{t-22} + \phi_{24}(o3d)_{t-24} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_4\varepsilon_{t-4}$$

Modelo 4 $$(o3d)_t = \phi_1(o3d)_{t-1} + \phi_2(o3d)_{t-2} + \phi_{22}(o3d)_{t-22} + \phi_{24}(o3d)_{t-24} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_4\varepsilon_{t-4}$$

Los modelos fueron estimados numéricamente arrojando los resultados que se muestran en la tabla 2.

De este conjunto de modelos se seleccionó el modelo 4 por presentar los mejores indicadores. La estimación de este modelo se presenta en la ecuación (3).

$$(o3d)_t = 0{,}529852(o3d)_{t-1} + 0{,}198980(o3d)_{t-2} + 0{,}075246(o3d)_{t-22} + \varepsilon_t + 0{,}395914\varepsilon_{t-1} - 0{,}066316\varepsilon_{t-4} \quad (3)$$

Diagnóstico

En esta sección se realizaron análisis de los residuales originados por el modelo (tabla 3). Inicialmente se observa que todos los coeficientes del modelo son significativamente diferentes a cero.

Posteriormente se verificó que las raíces de los polinomios característicos de la parte autorregresiva y de la parte de media móvil están todas dentro de círculo unitario, garantizando la estacionalidad e invertibilidad del modelo (figura 3).

Tabla 2 Posibles modelos generadores de la serie y sus características. R^2 es el coeficiente de determinación ajustado, AIC el criterio de información de Akaike, y SC el criterio de información de Schwarz [6].

Variable-Coeficiente-(sig)	R^2_{aj}	*AIC*	*SC*	*Observaciones*
Modelo 1 AR(1) 0,559921 (0,0000) AR(2) 0,183294 (0,0451) AR(24) 0,108251 (0,0000) MA(1) 0,372178 (0,0004) MA(4) -0,072033 (0,0007)	0,7278	1,5370	1,5499	Las raíces del polinomio característico están dentro del círculo unitario. Las correlaciones de los residuales no son significativas a un alfa del 5%
Modelo 2 AR(1) 0,777733 (0,0000) AR(24) 0,097196 (0,0000) MA(1) 0,157233 (0,0000) MA(4) -0,072732 (0,0009)	0,7273	1,5383	1,5486	Las raíces del polinomio característico están dentro del círculo unitario. El modelo es estacionario e invertible. Algunas de las correlaciones de los residuales son significativas a un alfa del 5%
Modelo 3 AR(1) 0,768539 (0,0000) AR(22) 0,066527 (0,0001) AR(24) 0,055727 (0,0012 MA(1) 0,160451 (0,0000) MA(4) -0,067178 (0,0022)	0,7292	1,5320	1,5449	Las raíces del polinomio característico están dentro del círculo unitario. El modelo es estacionario e invertible. Algunas de las correlaciones de los residuales son significativas a un alfa del 5%
Modelo 4 AR(1) 0,529852 (0,0000) AR(2) 0,198980 (0,0253) AR(22) 0,075246 (0,0001) AR(24) 0,065449 (0,0006) MA(1) 0,395914 (0,0001) MA(4) -0,066316 (0,0016)	0,7297	1,5304	1,5459	Las raíces del polinomio característico están por dentro del círculo unitario. El modelo es estacionario e invertible. Todas las correlaciones de los residuales son no significativas, los residuales son ruido blanco. Presenta la menor varianza de los residuales (0,269)

Tabla 3 Correlogramas residuales del modelo ARMA(24,4)

Retardo	*AC*	*PAC*	*Q-Stat*	*Prob*
1	-0,005	-0,005	0,0457	
2	-0,008	-0,008	0,1989	
3	0,010	0,010	0,4320	
4	-0,002	-0,002	0,4396	
5	0,003	0,003	0,4542	
6	-0,006	-0,006	0,5381	
7	-0,014	-0,014	0,9913	0,319
8	0,030	0,029	2,9241	0,232
9	-0,003	-0,003	2,9488	0,400

Tabla 3 (continuación)

Retardo	*AC*	*PAC*	*Q-Stat*	*Prob*
10	-0,013	-0,012	3,3320	0,504
11	0,001	0,000	3,3340	0,649
12	0,053	0,053	9,6255	0,141
13	0,010	0,010	9,8439	0,198
14	-0,014	-0,013	10,284	0,246
15	-0,007	-0,007	10,399	0,319
16	0,040	0,038	13,889	0,178
17	0,035	0,035	16,601	0,120
18	0,024	0,026	17,858	0,120
19	-0,006	-0,005	17,938	0,160
20	0,035	0,032	20,713	0,109
21	0.014	0,014	21,157	0,132
22	-0.018	-0,014	21,842	0,148
23	-0.013	-0,012	22,216	0,177
24	-0.013	-0,018	22,611	0,206

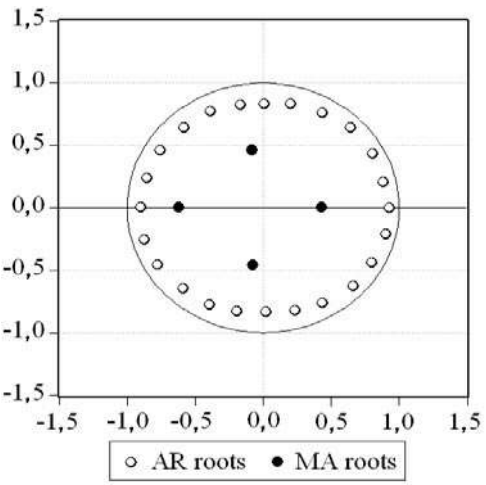

Figura 3 Raíces (AR: autorregresivas; MA: media móvil) de los polinomios característicos modelo 4: ARMA(24,4)

Los residuales originados por el modelo 4 no presentan autocorrelaciones simples o parciales significativas, lo cual indica que el modelo cumple con los supuestos fundamentales de estacionariedad requerida por la metodología Box-Jenkins, es decir, que los residuales se ajustan a un proceso de ruido blanco.

Validación

Con el fin de validar el modelo se utilizaron 264 datos horarios correspondientes a los 11 períodos de 24 horas finales de la serie, presentando un buen seguimiento de la serie por parte del modelo estimado. Para este procedimiento se trabajó con el lenguaje Matlab 7.

El modelo fue evaluado procesando las series proyectadas de 1 a 24 pasos hacia adelante, a las cuales se les calculó el error de pronóstico y se evaluó el modelo con base en estos resultados. La presencia de errores pequeños permite considerar este modelo como razonable.

Para implementar este proceso inicialmente se realizaron las transformaciones Box-Cox con $\lambda = -0{,}09$ *y* $c = 0$ y la desestacionalización con parámetros. La serie resultante se utilizó para realizar los pronósticos, también de 1 a 24 pasos adelante. Finalmente, a las series obtenidas se les aplicaron las transformaciones inversas a las utilizadas inicialmente, con el fin de llevar a cabo las comparaciones con lo valores reales.

En la evaluación de las series de pronósticos se tienen en cuenta los siguientes indicadores [7]:

1. Raíz del error cuadrático medio:

$$rms = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(o3p_t - o3o_t)^2}$$

2. Coeficiente de desigualdad de Theil:

$$U = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(o3p_t - o3o_t)^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}o3p_t^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}o3o_t^2}}$$

Donde:

$(o3o)_{T+j}$: es el valor observado para de la serie en T + j

$(o3p)_T(j)$: es el pronóstico realizado en T + j

Valores cercanos a cero indican un buen desempeño predictivo del modelo. Los resultados obtenidos se presentan en la figura 4. Se puede estimar que el poder de predicción es válido para los primeros 8 pasos (horas).

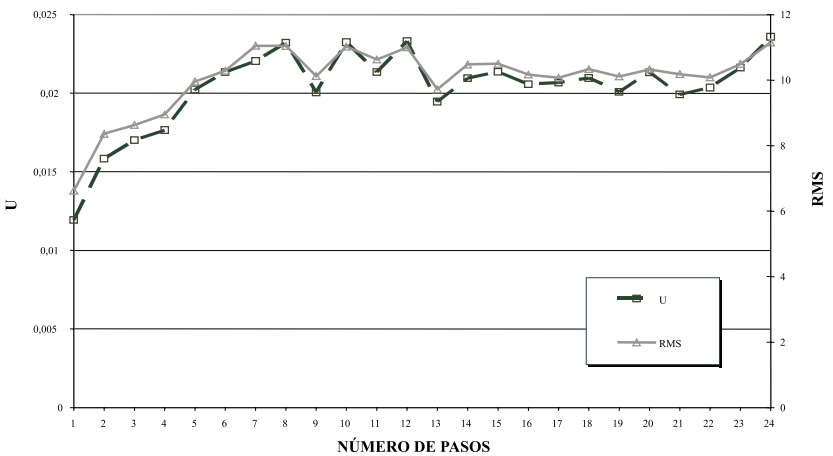

Figura 4 Errores de validación —Raíz del error cuadrático medio (RMS) y coeficiente de desigualdad de Theil (U) en función del número de pasos (horas)—

El siguiente modelo fue el elegido y se estimó con los 2.496 datos horarios iniciales:

$(o3d)_t$=	0,561706 $(o3d)_{t-1}$	+ 0,175131$(o3d)_{t-2}$	+ 0,063130$(o3d)_{t-22}$	+ 0,060175$(o3d)_{t-24}$
Valor-p	(0,0000)	(0,0478)	(0,0003)	(0,0008)
	+ 0,372373 e_{t-1}	- 0,059465 e_{t-4}		
Valor-p	(0,0003)	(0,0029)		

Los indicadores para este modelo son:

Coeficiente de determinación ajustado:

$R^2_{aj} = 0{,}724228.$

Criterio de información de Akaike IC :1,540681.

Criterio de información de Schwar SC :1,554790.

Conclusiones

El modelo que más se aproxima al proceso generador de la serie corresponde a un ARMA(24,4), cuya validación arroja un coeficiente de información ajustado R^2 de 0,72, que puede interpretarse en forma muy general como un buen nivel de concordancia entre el modelo y los datos.

El poder de predicción del modelo es aceptable hasta para un período de tiempo de 8 horas hacia adelante, como se puede apreciar en la figura 4. Se establece así la posible utilidad de esta metodología, mediante su implementación en tiempo real como complemento al análisis de los datos de monitoreo de calidad del aire, en la mitigación de los efectos negativos para la salud como consecuencia de episodios de altas concentraciones de ozono. En estos casos las medidas preventivas pueden incluir restricciones vehiculares y advertencias de exposición a sectores vulnerables como ancianos y niños, de forma que eviten el aire de exteriores en horas de máximas concentraciones de ozono.

El análisis univariado presentado en este documento constituye una herramienta simple comparada con otras técnicas multivariadas de estimación, las cuales generan estimaciones con calidades muy semejantes a las obtenidas en este trabajo.

Se requiere seguir realizando trabajos relacionados que permitan avanzar en el entendimiento de estos fenómenos y ayuden a las autoridades a generar directrices encaminadas a la conservación y mejoramiento de la calidad del aire.

Agradecimientos

Los autores agradecen al Departamento Administrativo de Gestión del Medio Ambiente, (DAGMA) por el suministro de los datos de monitoreo de calidad del aire.

Referencias

1. J. H. Lucio. *Desarrollo de modelos estocásticos lineales univariantes y multivariantes para la comprensión y predicción del ozono troposférico en atmósfera urbana*. Tesis doctoral, Universidad de Valladolid: Valladolid. 2003. pp. 65-82.
2. J. C. García. *Predicción del máximo de ozono utilizando metodología ARIMA sobre los datos monitorizados de calidad de aire de Valladolid.* Servicio del medio, Ayuntamiento de Valladolid: Valladolid 2003. pp. 1-8
3. G. E. P. Box, G. M. Jenkins, G. C. Reinsel. *Time Series Analysis. Forecasting and Control*: New Jersey Prentice Hall. 3.rd edition.1994. pp. 55-98
4. C. Chatfield. *The Analysis of Time Series: An Introduction.* 6.th ed. Bath. Chapman & Hall/CRC. 2004. pp. 46-262.
5. R.W. Boubel, D. Fox, D.B Turner. *Fundamentals of Air Pollution.* 3.th ed. Washington. Academic Press. 1994. pp. 165-178
6. W. Enders. *Applied Econometric Time Series.* 2.nd ed. Alabama. Wiley. 2004. pp. 52-68.
7. R. S. Pindyck, D. L. Rubinfeld. *Econometría: Modelos y Pronósticos*. México. McGraw-Hill Interamericana. 2001. pp. 404-410.