Seleccionar modelos de pronóstico para series de tiempo de contaminatnes de PM10 por criterio de Akaike y bayesiano

Alberto Benavides

Nuevo León, México

Abstract

La selección de modelos de pronóstico por criterios basados en similitud y número de parámetros es una estrategia más robusta para elegir parámetros de entrada. Además de dichos modelos, se parte de estrategias tentativas en la predicción de la serie de tiempo de los contaminantes del aire de PM10 en el Área Metropolitana de Nuevo León, México durante 2017.

Keywords: series de tiempo, AM, MA, ARMA, modelos de selección de criterios, Akaike, Bayes, AIC, BIC, PM10, Monterrey

1. Introducción

La capacidad de pronosticar acertadamente es una de las habilidades más valoradas en muchos de los ámbitos humanos que incluso aparece elogiado desde narraciones bíblicas y otras provenientes del periodo griego clásico [1]. La certeza de estos pronósticos ha sido relevante en la prevención de desastres naturales [2], el tratamiento preventivo de determinadas enfermedades (principalmente cáncer [3, 4]) o la elección de estrategias ventajosas en operaciones bursátiles [5, 6].

En los modelos de pronóstico es común utilizar estrategias intuitivas para elegir los parámetros con los que se harán las predicciones, sin embargo estas aproximaciones pueden considerarse poco formales como lo fueron las artes adivinatorias o proféticas representadas popularmente por Nostradamus [7].

En este artículo se pretenden utilizar criterios robustos para realizar la selección de parámetros en el pronóstico de contaminantes en el Área Metropolitana de Nuevo León, México (AMM). Para ello, se abordan en la sección 2 los fundamentos relacionados a las series de tiempo, sus características, modelos de pronóstico y los criterios con los que hará la selección de sus parámetros. Posteriormente, en la sección 3 se describen los datos, su origen, manipulación y preprocesamiento. Luego, se muestra en la sección 4 la metodología a la que se sometieron y los resultados obtenidos, por último, en la sección 5, se muestran las conclusiones obtenidas.

2. Marco teórico

Las series de tiempo son conjuntos de observaciones tomadas a lo largo del tiempo sobre algún evento, formalmente definidas [8] a partir del concepto de una familia de variables aleatorias $Z(\omega, t)$, con un espacio muestral ω y un conjunto de índices temporales t, en las que para una determinada t, $Z(\omega, t)$ es una variable aleatoria, y para una ω dada, $Z(\omega, t)$ es una **serie de tiempo** que, por comodidad, se denomina Z_t.

De estas series de tiempo se puede calcular su media

$$\mu_t = E(Z_t), \tag{1}$$

y varianza

$$\sigma_t^2 = E(Z_t - \mu_t)^2; \tag{2}$$

a partir de estas, la covarianza entre dos tiempos t_1, t_2

$$\gamma(t_1, t_2) = E(Z_{t_1} - \mu_{t_1})E(Z_{t_2} - \mu_{t_2}), \tag{3}$$

y la correlación también entre dos tiempos t_1, t_2

$$\rho(t_1, t_2) = \frac{\gamma(t_1, t_2)}{\sigma_{t_1}\sigma_{t_2}}. \tag{4}$$

Esto permite definir la **función de autocorrelación** (ACF) ρ_k como la correlación que tiene una serie consigo misma en los tiempos $t_1 = 0, t_2 = k$, es decir

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{5}$$

Ahora se introduce el concepto de **regresión lineal** entendida como la relación lineal entre la varialbe dependiente X_i y la variable independiente Y_i para $i = [1, \ldots, n]$, tal que

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{6}$$

Aquí, $\beta_0 + \beta_1 X_i$ son la función lineal que mejor se ajusta a X_i con base en el menor de los errores cuadrados, mientras que ϵ_i es la distancia entre dicha recta y el valor de Y_i para determinada i. Cuando se tienen p variables dependientes $X_{ij}, j = [1, \ldots, p]$, la regresión lineal se escribe

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i. \tag{7}$$

Con esto, la **función de autocorrelación parcial** (PACF) para un retraso de k unidades, queda definida a partir de la ecuación 7 como

$$Z_t = \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \ldots + \beta_k Z_{t-k} + \epsilon_t, \tag{8}$$

siendo el coeficiente β_k el que define la interacción del retraso k en la serie de tiempo Z_t.

Ahora bien, un **modelo autorregresivo** (AR) es uno de los modelos utilizados para el pronóstico de series de tiempo. Este modelo se basa en la idea de

que una serie de tiempo Z_t puede ser pronosticada Y_t a partir de la información proporcionada por las regresiones de momentos pasados de dicha serie de tiempo. A saber:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + +\phi_p Y_{t-p} + \epsilon_t \tag{9}$$

es el modelo autorregresivo de orden p, abreviado AR(p). En este tipo de modelos, es una práctica común utilizar los valores significativos de la PACF de una serie de tiempo como los valores p para obtener pronósticos.

Otro modelo utilizado para pronóstico es el **modelo de media móvil** en que en lugar de utilizar valores pasados de Z_t para realizar el pronóstico, como en los AR, se usan los errores de pronóstico ϵ_t de las predicciones con diferentes retrasos y coeficientes θ, por lo que

$$Y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}. \tag{10}$$

Estos dos modelos se combinan para formar el **modelo autorregresivo de media móvil** (ARMA) que tiene la forma

$$Y_t = c+\phi_1 Y_{t-1}+\phi_2 Y_{t-2}+\ldots+\phi_p Y_{t-p}+\epsilon_t+\theta_1 \epsilon_{t-1}+\theta_2 \epsilon_{t-2}+\ldots+\theta_q \epsilon_{t-q}, \tag{11}$$

denominado ARMA(p, q).

También, una serie de tiempo se puede entender a partir de su descomposición en la *tendencia* T_t, los *residuales* R_t y la **estacionalidad** S_t [9]. La tendencia es equivalente a β_0 y los residuales a los ϵ_t, mientras que la estacionalidad explica los patrones o ciclos que se hallan en las serie. Estas componentes pueden encontrarse sumadas

$$Z_t = T_t + R_t + S_t \tag{12}$$

o bien, multiplicadas:

$$Z_t = T_t \times R_t \times S_t. \tag{13}$$

Los modelos de pronóstico basados en el AR y MA son eficientes cuando se parte de una serie de tiempo **estacionaria**, esto es, aquella que tiene una media μ_t y varianza σ_t^2 constantes y no es estacional. Para determinar si una serie de tiempo es estacionaria, se cuenta con las pruebas de hipótesis de Dickey-Fuller aumentada (ADF) y la de Kwiatkowski-Phillips-Schmidt-Shinn (KPSS).

La hipótesis nula de la prueba ADF es que la serie es no estacionaria con un valor $p = 0.05$. Cuando no se puede rechazar la hipótesis nula, es posible hacer una diferenciación Z'_t de la serie de tiempo Z_t mediante

$$Z'_t = Z_t - Z_{t-d} \tag{14}$$

donde d es el retraso dado. Al realizar este proceso, el modelo ARMA se llama ARIMA, donde la I viene de *integrated* en inglés, que puede traducirse como diferenciada en este contexto. Se dice que se aplica un modelo ARIMA(p, d, q) a partir de un modelo AR(p), MA(q) con una serie de tiempo diferenciada d unidades.

A partir de la ACF, PACF de una serie de tiempo se pueden elegir las variables p, f de un modelo ARMA o ARIMA a partir de los valores estadísticamente significativos de dichas series. Las combinaciones de estos valores pueden ser variadas e incluir más o menos parámetros en los modelos o más o menos exactitud respecto a la serie que se desea pronosticar, por lo que se utilizan algunos criterios para determinar cuáles son las mejores combinaciones de parámetros para este tipo de pronósticos. Los dos modelos más usados son el Akaike (AIC) y el bayesiano (BIC).

El criterio de información de Akaike se describe como

$$\mathrm{AIC} = -2\log L + 2(p+q+d+1), \tag{15}$$

donde L es la similitud (definida en la ecuación 7.7.2 de [1]) entre la series de tiempo Z_t y el modelo Y_t. Por último, el criterio de información bayesiano [10] en series de tiempo depende del AIC, pero también toma en cuenta el número de muestras n en el modelo, así que se puede escribir

$$\mathrm{BIC} = -2\log(L) + \ln(n)\cdot(p+q+d+1). \tag{16}$$

Para estos criterios, es preferible un valor pequeño respecto a otro mayor porque esto implica el uso de menos parámetros $(p+q+d+1)$ y una mayor semejanza $2\log L$ entre serie de tiempo y función pronosticada.

3. Datos

La serie de tiempo que se desea pronosticar proviene de los registros de calidad de aire obtenidos del Sistema Integral de Monitoreo Ambiental de Nuevo León (México) [11] que cuenta con trece estaciones de monitoreo ambiental (cuya ubicación se muestra en la figura 1, p. 5) que registran fecha y hora, estación meteorológica, presión atmosférica, precipitaciones, humedad, radiación solar, temperatura, velocidad, dirección del viento y los contaminantes CO, NO, NO_2, O_3, SO_2, PM10, PM2.5. Algunas de estas estaciones registran datos desde 1993, y sólo coinciden su operación a partir de 2017, como se muestra en la figura 2 (p. 5).

Se llaman contaminantes PM10 a los que tienen que ver con partículas suspendidas con tamaño menor o igual a 10μm. Las altas concentraciones de estos contaminantes están relacionadas con enfermedades respiratorias [12] y muertes prematuras en población de riesgo [13].

Su estudio resulta interesante porque porque la Secretaría de Gobernación de México publicó la norma Norma NOM-172-SEMARNAT-2019 [14] en la que determina las concentraciones en las que el contaminante PM10 se consideran malas, mismas que se hallan en el cuadro 1 (p. 6). Al extraer los datos de PM10 durante el 2017 en el AMM, se comprueba que al menos un 65 % de los registros tienen una calidad considerada mala por la Secretaría de Gobernación de México, lo cual puede constatarse en la figura 3 (p. 6).

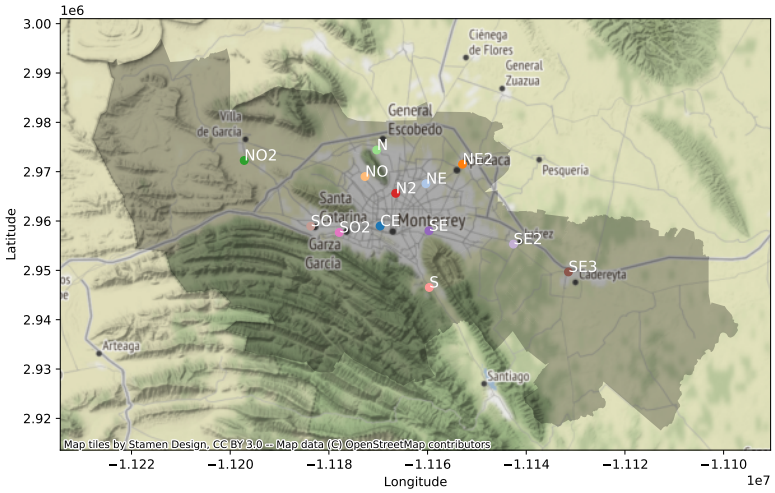

Figura 1: Trece estaciones de monitoreo del Área Metropolitana de Monterrey (Nuevo León, México)

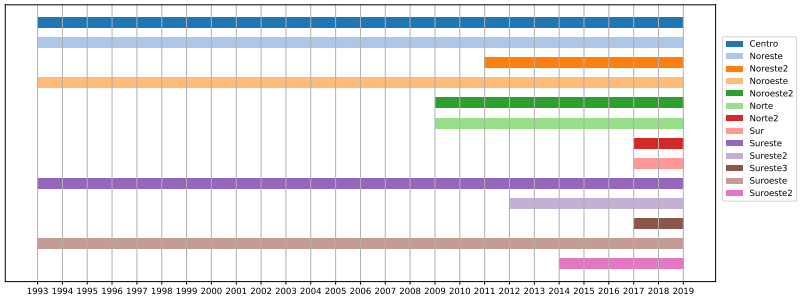

Figura 2: Barras de tiempo en que han estado activas las estaciones de monitoreo del AMM.

Tabla 1: Índice de aire y salud para PM10.

Calidad del aire	**Nivel de riesgo**	**12 horas (μg / m^3)**
Buena	Bajo	$[0, 50)$
Aceptable	Moderado	$[50, 75)$
Mala	Alto	$[75, 155)]$
Muy mala	Muy alto	$[155, 235)]$
Extremadamente mala	Extremadamente alto	> 235

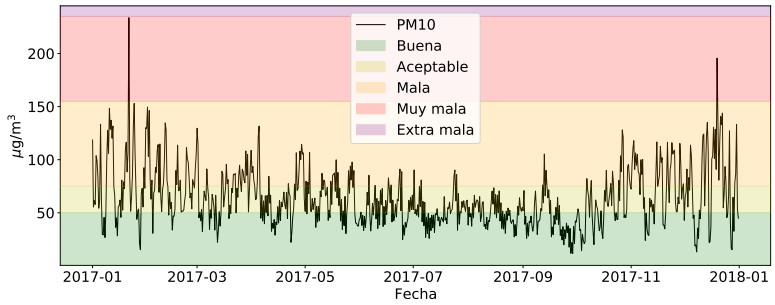

Figura 3: Serie de tiempo de PM10, durante el año 2017 para el AMM.

4. Metodología y resultados

La serie de tiempo se descompuso en las componentes de tendencia, estacionalidad y residuales, como se muestra en la figura 4 (p. 7). En estas imágenes, se puede observar que la tendencia β_0 no se mantiene constante, por lo que se realiza la prueba de Dickey-Fuller aumentada y se obtiene un valor $p = 0.51$, por lo que no se puede rechazar la hipótesis de que la serie no es no estacionaria, de modo que se debe hacer una diferenciación de la misma aplicando la ecuación 14. La serie tuvo que ser diferenciada dos veces, por lo que $d = 2$. La serie de tiempo original y la diferenciada con $d = 2$ están plasmadas en la figura 5.

También se obtuvieron sus ACF y PACF, incluidas en las figuras 6 y 6 (p. 8). En este caso, no hay manera de seleccionar por ninguna de las funciones de correlación un buen conjunto de parámetros, por lo que se procederá a generar modelos ARIMA$(p, 2, q)$ con $p = [1, \ldots, 10], q = [1, \ldots, 10]$, de los que se calcula el AIC y BIC, además de la suma de ambos y luego se muestran los diez menores valores de AIC + BIC y sus configuraciones de p y q en la tabla 2 (p. 8). En estos datos, se ve que la mejor combinación de valores es el modelo ARIMA$(2, 2, 5)$.

5. Conclusiones

Las dificultades que los modelos de pronóstico, como el ARIMA, presentan es que el tanteo de los parámetros del modelo de pronóstico pueden no conseguirse intuitiva ni directamente a partir de las ACF, PACF y diferenciación necesaria

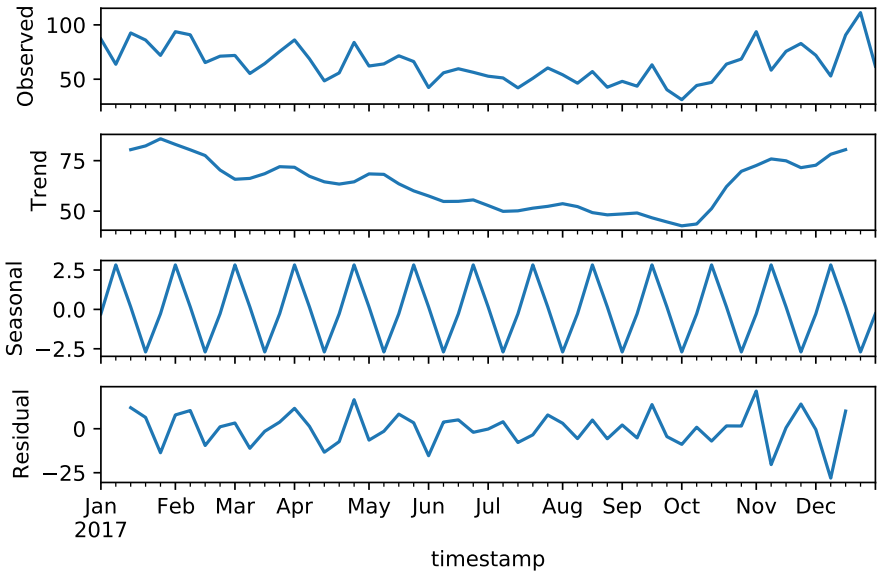

Figura 4: Serie de tiempo de PM10 de 2017 descompuesta en tendencia, estacionalidad y residuales.

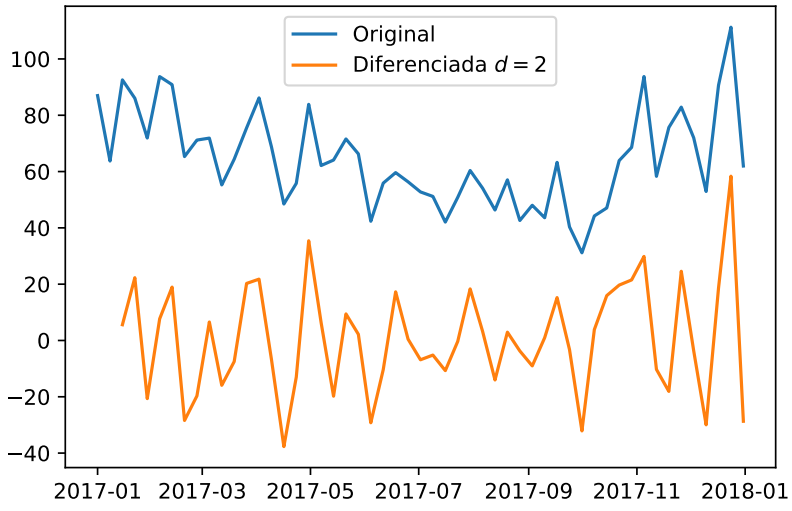

Figura 5: Serie de tiempo de PM10 de 2017 (azul) y la diferenciada en $d = 2$ (naranja).

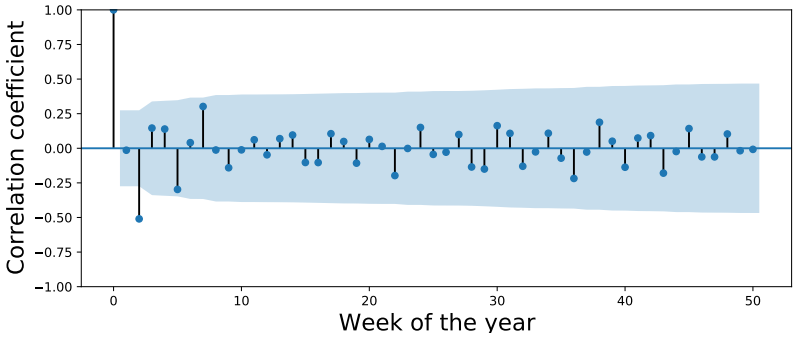

Figura 6: Coeficientes de autocorrelación para PM10 durante 2017 en el AMM.

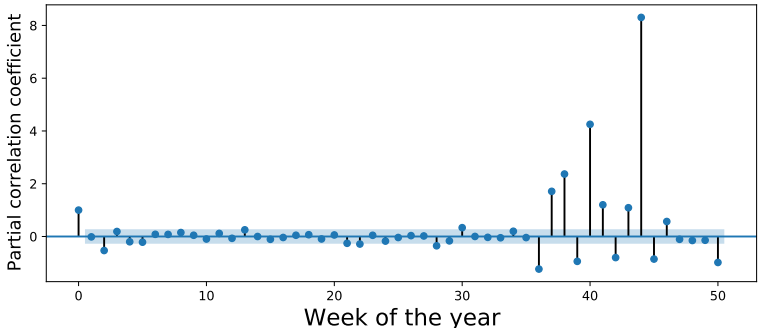

Figura 7: Coeficientes de autocorrelación parciales para PM10 durante 2017 en el AMM.

Tabla 2: Diez mejores combinaciones de p y q para ARIMA con base en la suma de AIC + BIC.

p	q	AIC	BIC	AIC + BIC
2	5	422.78	439.80	862.58
5	2	424.32	441.35	865.66
4	3	424.36	441.38	865.74
2	3	427.56	440.80	868.36
1	4	427.60	440.84	868.45
6	6	421.36	447.85	869.21
5	6	423.18	447.77	870.95
2	4	428.02	443.16	871.18
6	1	430.17	447.20	877.37
5	1	431.18	446.32	877.50

por prueba de estacionaridad. Los criterios de selección de modelos de Akaike y bayesiano resultan una herramienta estadística eficiente para solventar estas limitaciones. En general, sin embargo, es recomendable seguir la metodología descrita para tener una mejor información de la serie de tiempo que se desee pronosticar, pues desde la descomposición se tiene información relevante en la planeación de los modelos y estrategias a seguir en el pronóstico.

Referencias

[1] R. Hyndman, Forecasting : principles and practice, OTexts, Melbourne, 2018.

[2] H.-W. Cheng, Arima models for forecasting poisson process observations: Application to the volcanoes worldwide (2007). doi:10.25669/1RU9-PIWG.

[3] A. Earnest, S. M. Evans, F. Sampurno, J. Millar, Forecasting annual incidence and mortality rate for prostate cancer in australia until 2022 using autoregressive integrated moving average (ARIMA) models, BMJ Open 9 (8) (2019) e031331. doi:10.1136/bmjopen-2019-031331.

[4] N. Kumar, P. Kumari, P. Ranjan, A. Vaish, ARIMA model based breast cancer detection and classification through image processing, in: 2014 Students Conference on Engineering and Systems, IEEE, 2014. doi:10.1109/sces.2014.6880070.

[5] B. A. R. SK, Exchange rate forecasting using ARIMA, neural network and fuzzy neuron, Journal of Stock & Forex Trading 04 (03) (2015). doi:10.4172/2168-9458.1000155.

[6] Y. Xiao, J. Xiao, J. Liu, S. Wang, A multiscale modeling approach incorporating ARIMA and anns for financial market volatility forecasting, Journal of Systems Science and Complexity 27 (1) (2014) 225–236. doi:10.1007/s11424-014-3305-4.

[7] R. H. Popkin, Predicting, prophecying, divining and foretelling from nostradamus to hume, History of European Ideas 5 (2) (1984) 117–135. doi:10.1016/0191-6599(84)90063-9.

[8] W. Wei, Time series analysis univariate and multivariate methods, Pearson, Boston, 2019.

[9] P. Brockwell, Introduction to time series and forecasting, Springer, New York, 2002.

[10] Introduction to bayesian thinking, in: Bayesian Computation with R, Springer New York, pp. 19–37. doi:10.1007/978-0-387-71385-4_2.

[11] SISTEMA INTEGRAL DE MONITOREO AMBIENTAL [SIMA] NUEVO LEÓN, aire.nl.gob.mx — Home, http://aire.nl.gob.mx/, [Accesed 14/may/2020] (2020).

[12] C. A. Pope, D. W. Dockery, J. D. Spengler, M. E. Raizenne, Respiratory health and PM10pollution: A daily time series analysis, American Review of Respiratory Disease 144 (3_pt_1) (1991) 668–674. doi:10.1164/ajrccm/144.3_pt_1.668.

[13] K. Ito, G. Thurston, Daily pm10/mortality associations: an investigations of at-risk subpopulations., Journal of exposure analysis and environmental epidemiology 61 (1996) 79–95.

[14] Secretaría de Gobernación de México, Norma Oficial Mexicana NOM-172-SEMARNAT-2019, Lineamientos para la obtención y comunicación del Índice de Calidad del Aire y Riesgos a la Salud., SEMARNAT.
URL https://www.dof.gob.mx/nota_detalle.php?codigo=5579387&fecha=20/11/2019