ASADES

Avances en Energías Renovables y Medio Ambiente
Vol. 7, Nº 1, 2003. Impreso en la Argentina. ISSN 0329-5184

EVALUACIÓN DE CONCENTRACION DE DIOXIDO DE NITROGENO EN SALTA CAPITAL : UN ANÁLISIS ESTADÍSTICO ESTRUCTURAL

Orlando Avila Blas[1], Haydée Musso[2], Graciela Avila[2], Analía Boemo[2] y Ramón Farfán[2]
[1,2]Av. Bolivia 5150 - (A4408FVY) Salta, Argentina, Tel. : (0387) 4255385, 4255354
Email : oblas@unsa.edu.ar , hmusso@unsa.edu.ar , gavila@unsa.edu.ar , aboemo@unsa.edu.ar , raf@unsa.edu.ar

RESUMEN- En el presente trabajo se modelan estadísticamente los datos de dos series de determinaciones de dióxido de nitrógeno en la atmósfera baja. Se utiliza la metodología de estudio estructural, desarrollándose la teoría matemática estadística específica correspondiente, a fin de presentar los dos modelos que se pueden utilizar para generar sintéticamente valores de la variable de estudio. Esta modelización permite realizar pronósticos con muy alta confiabilidad (95%).

Palabras claves : muestreadores pasivos, contaminación del aire, estadística, modelado estructural, pronósticos.

INTRODUCCIÓN

La concentración de óxidos de nitrógeno en la atmósfera baja de una región determina la calidad del aire que hay en ella. Los óxidos de nitrógeno, y particularmente el dióxido de nitrógeno, pueden impactar negativamente en la salud humana y contribuir a la degradación del medio ambiente por generación de lluvia ácida y formación del nebluno fotoquímico.

El 90% de los óxidos de nitrógeno producidos por el hombre se originan en las combustiones. Existen dos fuentes de nitrógeno que contribuyen a la formación de sus óxidos en las reacciones de combustión, el comburente aire que contiene nitrógeno y oxigeno molecular en una relación molar aproximada de 4 a 1 respectivamente, y los combustibles que contienen nitrógeno en su composición como el gas-oil, fuel-oil y algunas naftas. El mayor porcentaje de las emisiones de NO_x tanto de fuentes fijas como de fuentes móviles proceden del consumo de estos tipos de combustibles ya que el gas natural está esencialmente libre de compuestos nitrogenados. También los incineradores de residuos municipales son una fuente considerable de óxidos de nitrógeno debido a la combustión de grandes cantidades de desechos orgánicos. Un porcentaje mucho menor es el aportado por la descomposición de la materia orgánica. Las muestras de óxidos de nitrógeno fueron colectadas a través de muestreadores pasivos difusionales tipo Palmes, conteniendo como absorbente trietanolamina y un surfactante (Brij-35) (Gair, A.J., et al., 1991). Los muestreadores, colocados a dos metros de altura, fueron expuestos por triplicado durante un mes, en forma continua, desde enero de 2001 a octubre de 2002 para los dos sitios estudiados. El análisis de los NO_x absorbidos se realizó inmediatamente después de retirar los muestreadores, desarrollando la reacción de Saltzman para NO_2^- , y expresando la concentración de NO_x como NO_2 (Musso et al, 2002).

Area de estudio: los sitios de muestreo seleccionados para este trabajo se encuentran ubicados a unos 6 km al Sur-Este de la ciudad de Salta. El sitio Parque Industrial cuenta con cuarenta y cinco establecimientos fabriles de naturaleza variada; como ejemplo se puede mencionar las fábricas de detergentes y desinfectantes, de fertilizantes, de cigarrillos, de poliuretano, de baterías y placas para vehículos, de grasas y harina de carne, de derivados de papel, de muebles, de ropa, de alimentos, de vinagre, de cerámicos, una deshidratadora de ulexita, una curtiembre y empresas de fundición de hierro y acero, metalúrgicas, molienda de minerales, así como fraccionadoras de carbón. Al Sur-Este del Parque Industrial (aproximadamente a 1 km) se encuentra el segundo sitio de muestreo, ubicado entre el predio de la Planta de Tratamiento de Líquidos Cloacales y el enterramiento de residuos municipales San Javier. La Planta de Tratamiento de Líquidos Cloacales cuenta con una capacidad para satisfacer la demanda de 350.000 habitantes, mientras que San Javier compacta y tapa por día unas 300 toneladas de residuos municipales en un predio de 24 hectáreas, de las cuales 17 ya se encuentran ocupadas. Las 10 primeras hectáreas ocupadas que operaron desde su inauguración (1994) hasta 2000 no recibieron el actual tratamiento, produciéndose en esa zona incendios espontáneos como consecuencia de la descomposición de los residuos.

[1] Probabilidades y Estadística, Depto. de Matemática, Fac. Cs. Exactas – U.N.Sa-C.I.U.N.Sa. Proyecto Nº 1009
[2] Química Analítica, Depto. de Química, Fac. Cs. Exactas – U.N.Sa-C.I.U.N.Sa. Proyecto Nº 1009

El estudio de series de tiempo de variables físico-químicas comenzó formalmente en el intervalo de fines de los años '70 y comienzos de los '80; los investigadores comenzaron a utilizar la teoría de *modelización estadística* para generar series temporales (series sintéticas, indiferenciables de las reales) de valores diarios de irradiación. Últimamente se han propuesto métodos estadísticos de generación de series de algunas variables de interés : temperatura de superficie (Avila Blas, 1997), radiación horaria (Aguiar, 1992), utilizando modelos basados en el tratamiento de Box y Jenkins, es decir, modelos ARIMA(p,d,q). También se han conseguido avances significativos modelando series de radiación para el caso en que la varianza asociada a los disturbios o ruidos blancos de la serie pueda variar aleatoriamente y sin tener que imponer sobre los datos la condición de tener una distribución normal o gaussiana (hecho destacable ya que los datos asociados a radiación no poseen este tipo de distribución probabilística). El año pasado se encontraron modelos estructurales de series medidas en las ciudades de Córdoba, Marcos Juárez y Paraná (Avila Blas *et al.*, 2002) con muy buenas propiedades estadísticas, mediante la introducción de una transformación matemática especial denominada "índice de contribución relativa", que permitió "levantar" la condición de normalidad en la distribución de los datos y atacar la presencia de datos atípicos o "outliers". Pero hasta el momento, no se conocen tratamientos de este tipo aplicados a datos de concentración de contaminantes en el aire.

ANÁLISIS ESTADÍSTICO ESTRUCTURAL UTILIZADO.

La idea básica de los modelos estructurales de series de tiempo es que ellos pueden ser puestos como modelos de regresión en donde las variables explicativas son funciones del tiempo, con coeficientes que pueden cambiar a través del tiempo. La estimación actual de los coeficientes ó filtrada se logra poniendo al modelo en forma de espacio de estado y aplicándole luego el denominado Filtro de Kalman. Se emplean algoritmos específicos para hacer predicciones y para los suavizados. Esto último significa computar el mejor de los estimadores en todos los puntos de la muestra usando al conjunto de observaciones. La magnitud por la cual los parámetros pueden variar está gobernada por los llamados hiperparámetros. Estos pueden ser estimados por el método de máxima verosimilitud, construyendo la función específica a optimizar. El énfasis está puesto en la formulación del modelo en términos de componentes cuya presencia está sugerida por el conocimiento del fenómeno bajo estudio, de sus aplicaciones o por una inspección del gráfico de la serie original. Una vez que el modelo ha sido estimado, el mismo tipo de tests de diagnóstico para los modelos ARIMA puede ser aplicado. Además el estudio se completa con tests de falta de normalidad y heterocedasticidad (varianza variable en el tiempo), tests para la calidad predictiva en períodos posteriores a la muestra y gráficos de los componentes suavizados. Además no es necesario que la serie sea estacionaria, es decir, con media y varianza constante en el tiempo, lo que sucede muy a menudo en variables de polución.

El tratamiento estadístico de los modelos estructurales de series de tiempo está basado en la forma de espacio de estado, el filtro de Kalman y el suaviazador asociado. La función de verosimilitud se construye a partir del filtro de Kalman en términos de la predicción un paso hacia adelante, y se maximiza con respecto a los hiperparámetros por optimización numérica (Koopman *et al*., 1995). El vector marcador ("score") de los parámetros puede obtenerse a través de un algoritmo de suavizado asociado al filtro de Kalman. Una vez que los hiperparámetros fueron estimados, el filtro se usa para conseguir predicciones de los residuos un paso adelante, lo que nos permite calcular los estadísticos para probar normalidad, correlación serial y bondad de ajuste. El suavizador se usa para estimar los componentes que no son observables, como por ejemplo, tendencia y estacionalidad, y para el cálculo de estadísticos que son empleados para detectar observaciones atípicas ("outliers") y cambios estructurales. El enfoque de espacio de estado es particularmente interesante de ser empleado cuando la serie tiene datos faltantes o han sido incorporados temporalmente. Otra importancia de este enfoque es que los modelos estructurales pueden ser escritos mediante transformaciones adecuadas, como modelos ARIMA: ésta es la denominada forma reducida. La representación matemática de un modelo de espacio de estado relaciona al vector de disturbios $\{e_t\}$ con el vector de observaciones $\{y_t\}$ a través de un proceso de Markov $\{a_t\}$ y es :

$$\begin{aligned} y_t &= Z_t a_t + G_t e_t , \quad e_t \sim N(0, H_t) \\ a_t &= T_t a_{t-1} + H_t \eta_t , \quad \eta_t \sim N(0, Q_t) ; \quad a_0 \sim N(\boldsymbol{a_0}, \boldsymbol{P_0}) , \quad t=1,\ldots,T \end{aligned} \tag{1}$$

donde a_t es el vector de estado de orden mx1, e_t es un vector de disturbios de orden kx1 y las matrices del sistema Z_t, T_t, G_t, y H_t tienen dimensiones Nxm, mxm, Nxk y mxk respectivamente. Los disturbios son ruido blanco mutuamente no correlacionados con medio cero y varianza H_t. Cuando se supone normalidad, los disturbios son independientes entre sí. Las matrices G_t y H_t pueden interpretarse como matrices de selección, lo que le brinda generalidad al modelo. Las cuatro matrices son fijas y si en ellas hubiere elementos desconocidos, se incorporan al vector Ψ de hiperparámetros, el que es estimado por máxima verosimilitud. Los estadísticos de prueba usados para la bondad de ajuste son : el BS (de Bowman y Sentón), que emplea estimadores de la asimetría y de la kurtosis de los datos, el Q (de Box-Ljong) para la autocorrelación seriel y el de DW (de Durbin y Watson), los que junto con otros estadísticos adicionales, califican según sus valores, al modelo como adecuado para ajustar a los datos observados.

ANALÍSIS ESTADÍSTICO DE LAS SERIES

Las series originales tratadas aquí constan de T = 26 (serie 1: octubre de 2000 a noviembre de 2002, para la Planta Industrial) y T = 19 (serie 2: mayo de 2001 a noviembre de 2002, para la Planta de Tratamiento). El reducido número de datos no es impedimento para emplear el tratamiento estructural. La representación gráfica de las series (figuras 1 y 3), junto con sus

correspondientes correlogramas (figuras 2 y 4). Esta función es adimensional, por lo que no se fija una escala en el eje de las abscisas).

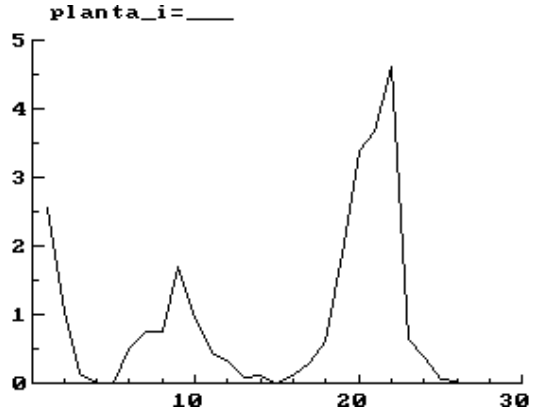

Figura 1 : Serie de NO_2 ($\mu g/m^3$), (eje *x* en meses)

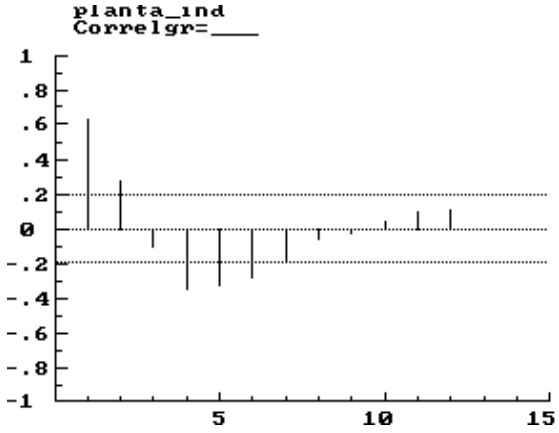

Figura 2 : correlograma de la serie bajo estudio

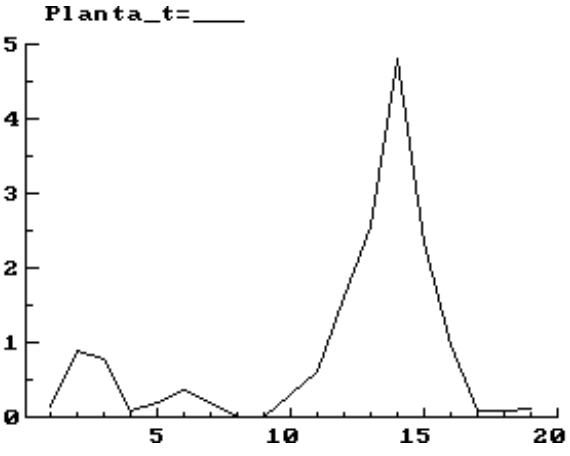

Figura 3 : Serie de NO_2 ($\mu g/m^3$), (eje *x* en meses)

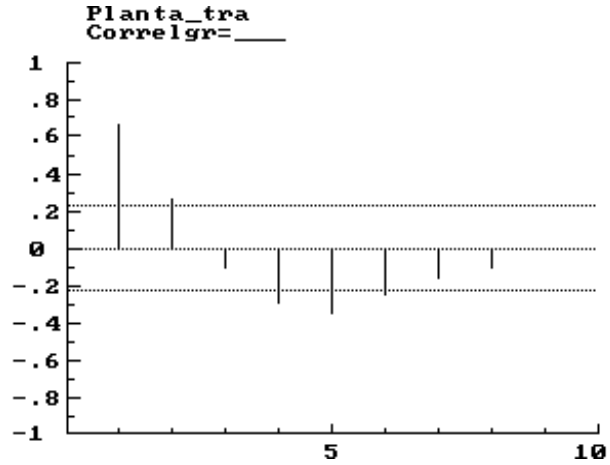

Figura 4 : correlograma de la serie bajo estudio

Las series respectivas tienen una distribución empírica con una pronunciada cola por la derecha ó "cola pesada" (asimetría positiva extrema), tal como se observa en las figuras 5 y 6 :

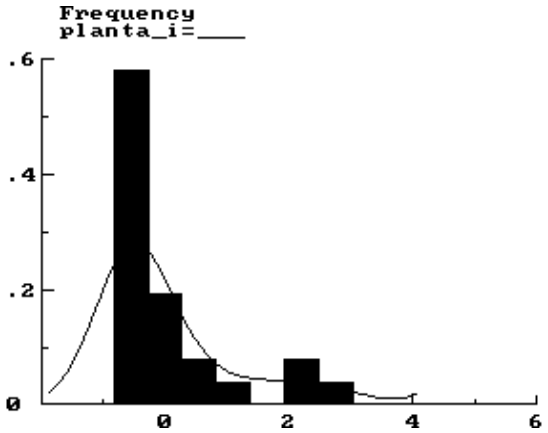

Figura 5 : Distribución para la serie 1, (eje *x* en meses)

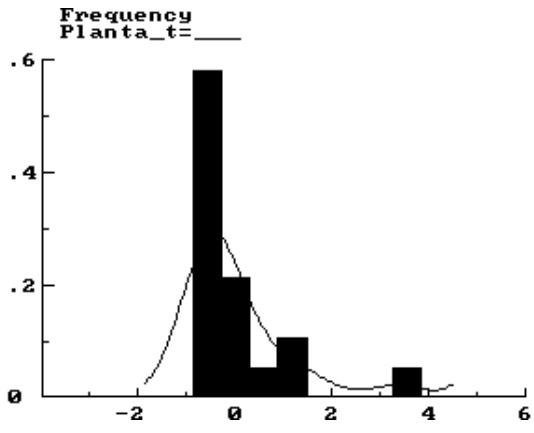

Figura 6 : Distribución para la serie 2, (eje *x* en meses)

Antes de llegar al modelo final, se han intentado diferentes modelos : tomar la variable original y considerarla como una combinación lineal de componentes de nivel e irregular estocásticas, ó bien, realizarle una diferenciación de orden 1 y proponer la misma descomposición anterior, y otros , pero en todos ellos el modelado no resultó con las propiedades deseadas, dado que los

valores de los estadísticos asociados a la bondad de ajuste (BS, Q y DW) no fueron los adecuados. Además, la observación complementaria de las funciones de distribución empírica (figuras 5 y 6 respectivamente) sugiere emplear una metodología diferente a las anteriormente nombradas. El hecho de observar una pesada cola a derecha de las distribuciones respectivas llevó a pensar la resolución del problema enmarcándola dentro de la familia de distribuciones exponenciales, en particular la distibución de Poisson. La transformación final que permitió modelar las series corresponde a un ***Modelo de Regresión Dinámica con observaciones de tipo Poisson***

En primer lugar, cabe recordar que la teoría general para el tratamiento de observaciones de la familia exponencial puede verse con más detalle en J. C. Abril, 1999, y está basada en técnicas iterativas para computar el EMP (error medio de los parámetros) empleando algoritmos rápidos del SFK (suavizado y filtrado de Kalman), y también métodos para calcular estimadores aproximados por máxima verosimilitud de los hiperparámetros involucrados. Para las series bajo estudio, se tomaron aspectos de esta teoría y se diseñó un modelo de regresión que tiene un solo regresor z_t con coeficiente α_t que varía en el tiempo de acuerdo a un camino aleatorio de la forma

$$\alpha_t = \alpha_{t-1} + \eta_t \quad \text{con} \quad \eta_t \sim N(0, \sigma^2) \;;\; t=1,2....T \tag{2}$$

Las observaciones y_t son variables con distribución de Poisson con medias $Exp(z_t \alpha_t)$. Esta es una forma simple de modelo loglineal y es un caso particular de un modelo más general, con la diferencia de que aquí el coeficiente de regresión varía en el tiempo. Sin pérdida de generalidad, podemos tomar para compensar el proceso de inicialización, al vector de estado inicial α_0 como una constante conocida, y en consecuencia, el logaritmo de la función de densidad conjunta de α_t e Y_n , salvo algunas constantes, es

$$\log\, p(\alpha, Y_n) = \frac{-1}{2\sigma^2}\sum_{t=1}^{T}(\alpha_t - \alpha_{t-1})^2 + \sum_{t=1}^{T}[z_t\alpha_t y_t - \exp(z_t\alpha_{t-1}) - \log(y_t)!] \tag{3}$$

Diferenciando con respecto a a e igualando a cero, obtenemos los estimadores EMP para $a_1,...,a_T$ como solución de las ecuaciones

$$\frac{1}{\sigma^2}(\alpha_{t-1} - 2\alpha_t + \alpha_{t+1}) + z_t[y_t - \exp(z_t\alpha_t)] = 0, t = 1,2,..., T-1 \tag{4}$$

con el primer término reemplazado por : $\sigma^{-2}(\alpha_{n-1} - \alpha_n)$, para t=T (5)

Estas son ecuaciones de tipo no lineal y para poder resolverlas se procede a considerar el modelo gaussiano análogo. Este tiene el mismo camino aleatorio que el dado por la ecuación (2), pero en lugar del modelo loglineal, tenemos el modelo de regresión lineal

$$y_t = z_t\alpha_t + \varepsilon_t \;, \text{con } \varepsilon \sim N(0, \sigma_t^2) \;;\quad t=1,2....T \tag{6}$$

el cual es un modelo gaussiano de regresión por el origen, con un coeficiente que varía en el tiempo y con disturbios heterocedásticos. El logaritmo de la función de densidad conjunta viene dado por

$$\log p(\alpha, Y_n) = \frac{-1}{2\sigma^2}\sum_{t=1}^{T}(\alpha_t - \alpha_{t-1})^2 + \sum_{t=1}^{T}(y_t - z_t\alpha_t)^2 \tag{7}$$

el cual, diferenciando e igualando a cero, nos da para el EMP, las ecuaciones

$$\frac{1}{\sigma^2}(\alpha_{t-1} - 2\alpha_t + \alpha_{t+1}) + \frac{z_t}{\sigma_t^2}[y_t - z_t\alpha_t] = 0, t = 1,2,..., T-1 \qquad \text{con } \sigma^{-2}(\alpha_{n-1} - \alpha_n) \text{ para t=T} \tag{8}$$

A fin de poder resolver las ecuaciones dadas por (4), las linealizamos poniéndolas de la misma forma que en (8), para luego resolver las ecuaciones resultantes mediante el SFK. Supongamos ahora que $\overset{>}{\alpha_t}$ es un valor experimental de $\hat{\alpha_t}$; entonces, si expandimos alrededor del valor dado, obtenemos:

$$\hat{y}_t = y_t - \exp(z_t \hat{\alpha}_t)(1 - z_t \hat{\alpha}_t),\ \hat{z}_t = \exp(z_t \hat{\alpha}_t) z_t,\ \hat{\sigma}^2_t = \exp(z_t \hat{\alpha}_t) \tag{9}$$

De este modo el segundo sumando en la derecha de (4) se convierte en: $\frac{\hat{z}_t}{\hat{\sigma}^2_t}(\hat{y}_t - \hat{z}_t \hat{\alpha}_t)$ (10)

el que tiene la misma forma que el segundo término de (8), y en consecuencia podemos resolver la ecuación linealizada, y a partir de ésta obtener una mejor aproximación a los estimadores $\hat{\alpha}_t$ para el modelo Poisson, aplicando exactamente el mismo algoritmo del SFK que se usa para la estimación de los correspondientes al modelo gaussiano (2) y (6). Los valores resultantes son sustituídos en (8) para obtener nuevos valores de los $\hat{y}_t$, $\hat{z}_t$ y $\hat{\sigma}^2_t$, los que a su vez son tratados con el SFK y así sucesivamente hasta lograr la convergencia deseada. Cualquier método de inicialización considerado en los modelos gaussianos pueden ser empleados de ahora en adelante. Esta metodología se aplicó a las series en cuestión, tomando como criterio de convergencia para el vector de estado, parar cuando el cambio relativo promedio en el EMP para todos los elementos del estado para t = 1,2,...,T era menor que 10^{-6}. El criterio para realizar la estimación de los hiperparámetros fue el de parar el proceso iterativo cuando el cambio relativo promedio de los hiperparámetros fuese menor que 10^{-5}, habiendo iniciado el proceso de estimación de los hiperparámetros por el método del filtro de Kalman extendido. El proceso recursivo conduce a la definición de una variable transformada final, denominada "Svar22" (definida por la ecuación 9). Las resultados computacionales más importantes obtenidos se muestran en la tabla 1 , usando el Soft STAMP 5.0 (Koopman et al, 1995) con respecto al modelado de las variables bajo estudio, y para el caso de las serie correspondiente a la planta industrial (los de la planta de tratamiento son similares).

Ecuación final	Resumen de estadísticos
	Svar2
Svar2= Nivel fijo + AR(1) + 1 Ciclo + Irregular	Std.Error 5.918
	Normalidad 2.167e+005
Desviaciones estándar estimadas de los disturbios	H(239) 0.39937
	r(1) 0.0012299
Componente Svar22 (q-ratio)	r(25) 0.045725
Irr 5.7339 (1.0000)	DW 1.994
Cy1 0.00000 (0.0000)	Q(25,20) 36.52
Ar1 0.392968 (0.0685)	R^2 0.80424

Tabla 1. Resultados estadísticos más destacadas correspondientes al modelo final

En la figuras siguientes, y para el caso de la serie de la planta industrial, se muestran las series original (línea continua) versus la serie modelada (línea débil), indicando el proceso de pronóstico comparativo para los 5 períodos, así como también el comportamiento residual (correlograma controlado)

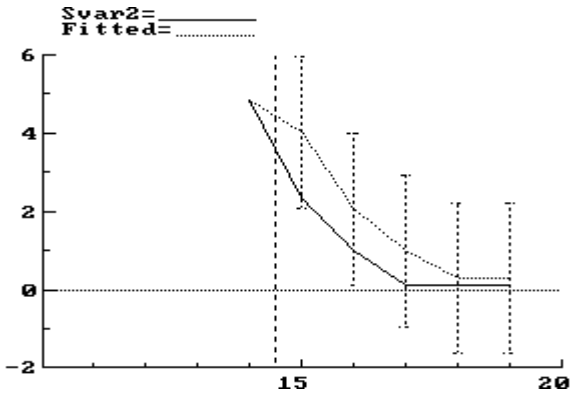

Figura 7: pronósticos comparativo de los últimos 5 meses

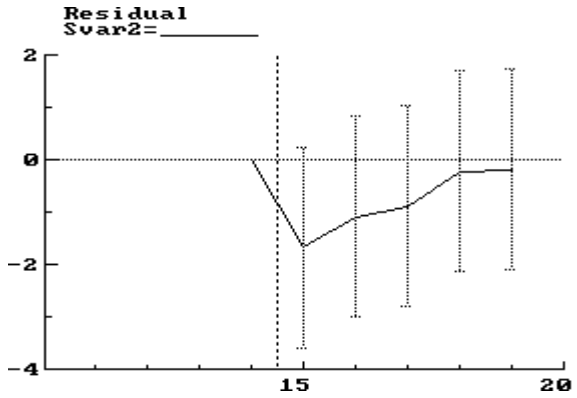

Figura 8: comportamiento residual de la serie modelada

CONCLUSIÓN

Mediante el empleo del tratamiento estructural de series de tiempo y empleando una estructura matemática basada en la distribución de Poisson, se consiguió modelar dos series de valores de concentración de dióxido de nitrógeno, con una bondad de ajuste muy buena. Esto permite realizar estimaciones a corto plazo con una confiabilidad alta, del 95%.

REFERENCIAS

Abril J. C. (1999). *Análisis Estadístico de Series de Tiempo Basado en Modelos de Espacio de Estado*. E.U.D.E.B.A., 1999.

Aguiar R. y Collares Pereira M. (1992). Tag : A time dependent, autorregressive, gaussian model for generating synthetic hourly radiation, *Solar Energy*, **49**, 3, 167-174.

Avila Blas O. J. (1997). Análisis Espectral de Series de Temperatura de Superficie. *Revista FACENA, Univ. Nac. Nordeste*, **13**, 79-99.

Avila Blas O.J. y Grossi Gallegos H. (2002). Modelos estadísticos estructurales de series de irradiación solar global diaria para Córdoba, Marcos Juarez y Paraná. *Avances en Energías Renovables y Medio Ambiente,* **6,** 11.07-11.11

Gair, A.J., Penkett S. A. And Oyola, P., 1991. Development of a simple passive technique for the determination of nitrogen dioxide in remote continental locations. *Atmospheric Environment* , **25A**, 9: 1927-1939

Koopman S. J., Harvey A. C., Doornik J. A. y Shepard N. (1995). *STAMP 5.0, Structural Time Series Analyser, Modeller and Predictor.* 1a. edición. Chapman and Hall, London.

Musso H., Boemo A., Avila G., Farfán R.(2002). Concentraciones de Ozono y de Dióxido de nitrógeno en la troposfera de Salta (Capital) . *Avances en Energías Renovables y Medio Ambiente,* **6,** 01.17-01.22

ABSTRACT- In this paper two series of nitrogen dioxide concentration data in the low atmosphere are modelled. The method of structural analysis is used, and the specific mathematical statistic theory is presented for both models, which can be used for generating synthetic values for the variable under treatment. Forecasting with a high confidential level (95%) can be made with this modelling.

Keywords: passive samplers, air pollution, statistic, structural models, forecasting