Revista EIA, ISSN 1794-1237 Número 18, p. 19-31. Diciembre 2012
Escuela de Ingeniería de Antioquia, Medellín (Colombia)

AJUSTE DE CURVAS MEDIANTE MÉTODOS NO PARAMÉTRICOS PARA ESTUDIAR EL COMPORTAMIENTO DE CONTAMINACIÓN DEL AIRE POR MATERIAL PARTICULADO PM10

Jhovana Reina*,
Javier Olaya**

RESUMEN

Uno de los principales agentes contaminantes del aire es el material particulado de diámetro aerodinámico inferior a 10 micrómetros, comúnmente conocido como PM10. Su comportamiento varía de forma irregular y temporal en la atmósfera, debido a las actividades humanas, condiciones atmosféricas inestables y fenómenos meteorológicos. El propósito de este estudio es caracterizar con un modelo de suavización no paramétrica el comportamiento del PM10 en el aire a lo largo de un día, teniendo en cuenta el día de la semana y los niveles de precipitación. El modelo propuesto se ilustra con registros sobre contaminación por PM10 y con datos de precipitación en el norte de Cali, Colombia. Se estiman curvas típicas diarias del comportamiento del PM10 usando suavizadores *kernel* y *spline*. El procesamiento se ejecuta con el software estadístico de libre distribución R. Las curvas estimadas permiten observar un comportamiento unimodal del PM10 durante las horas de la mañana, diferenciado por días de la semana y por días con lluvia y sin lluvia. Los modelos permiten caracterizar de manera robusta el comportamiento diario del PM10, teniendo en cuenta observaciones heterocedásticas bajo un escenario de múltiples respuestas por punto de diseño.

PALABRAS CLAVE: contaminación atmosférica; heterocedasticidad; PM10; regresión no paramétrica; suavización *kernel*; suavización *spline*.

* Estadística, Universidad del Valle, Cali, Colombia. Joven investigadora Colciencias. jhoreina@univalle.edu.co

** Estadístico, Universidad del Valle; MSc y PhD in Mathematical Sciences, Clemson University. Profesor Titular, Universidad del Valle, Cali, Colombia. olaya@univalle.edu.co

Artículo recibido 20-III-2012. Aprobado 24-VI-2012
Discusión abierta hasta junio de 2013

CURVE FITTING NONPARAMETRIC METHODS FOR STUDYING BEHAVIOR FROM AIR POLLUTION PM10

ABSTRACT

One of the main air pollutants is the particulate matter whose aerodynamic diameter is less than 10 micrometers, usually referred as PM10. It is a fact that the PM10 behavior in the air varies in an irregular way, and also in a temporal way in the atmosphere, mainly due to human activities, to unstable atmospheric conditions, and to meteorological phenomena. Our main purpose is to characterize through a nonparametric smooth model the PM10 daily behavior, taking into account the day of the week, and the precipitation levels. We illustrate the model using records on PM10 contamination, as well as on data on rain precipitation in the north side of Cali, Colombia. We estimate daily typical curves of the PM10 behavior using kernel and spline estimators. We processed these data using the free distribution statistical software R. The estimated curves allow us to observe a PM10 unimodal behavior during the morning hours, which varies from one day to another and from rainy to non-rainy days. The fitted models allow a robust characterization of the PM10 daily behavior, considering heteroscedastic observations on a multiple response per design point scenario.

KEY WORDS: air pollution; heteroscedasticity; PM10; nonparametric regression; kernel smoothing; spline smoothing,

AJUSTE DE CURVAS MEDIANTE MÉTODOS NÃO PARAMÉTRICOS PARA ESTUDAR O COMPORTAMENTO DE CONTAMINAÇÃO DO AR POR MATERIAL PARTICULADO PM10

RESUMO

Um dos principais agentes contaminantes do ar é o material particulado de diâmetro aerodinâmico inferior a 10 micrômetros, comumente conhecido como PM10. Seu comportamento varia de forma irregular e temporal na atmosfera, devido às atividades humanas, condições atmosféricas instáveis e fenômenos meteorológicos. O propósito deste estudo é caracterizar com um modelo de suavização não paramétrica o comportamento do PM10 no ar ao longo de um dia, tendo em conta o dia da semana e os níveis de precipitação. O modelo proposto ilustra-se com registros sobre contaminação por PM10 e com dados de precipitação no norte de Cali, Colômbia. Estimam-se curvas típicas diárias do comportamento do PM10 usando suavizadores *kernel* e *spline*. O processamento executa-se com o software estatístico de livre distribuição R. As curvas estimadas permitem observar um comportamento unimodal do PM10 durante as horas da manhã, diferenciado por dias da semana e por dias com chuva e sem chuva. Os modelos permitem caracterizar de maneira robusta o comportamento diário do PM10, tendo em conta observações heterocedásticas baixo um cenário de múltiplas respostas por ponto de desenho.

PALAVRAS-CÓDIGO: contaminação atmosférica; heterocedasticidade; PM10; regressão não paramétrica; suavização *kernel*; suavização *spline*.

1. INTRODUCCIÓN

El material particulado es una mezcla de sustancias sólidas y líquidas suspendidas en el aire que, dependiendo de sus características físicas y químicas, pueden generar varios efectos nocivos en la salud de los seres humanos y en los ecosistemas ambientales. Un indicador de este tipo de contaminante se conoce como PM10, partículas cuyo diámetro es inferior a 10 micrómetros, que al penetrar la tráquea, los pulmones y los bronquios pueden causar múltiples enfermedades tales como afecciones pulmonares, asma, obstrucción pulmonar crónica y cáncer pulmonar. También son fuente potencial de problemas en la vista, problemas cardiovasculares y congestiones cardiacas (Perez-Padilla, Schilmann y Riojas-Rodriguez, 2010).

Las actividades humanas, por ejemplo, la industria y el movimiento vehicular, se consideran las principales fuentes de emisión de material particulado que, en conjunto con factores meteorológicos y topográficos, hacen que el comportamiento del PM10 varíe temporalmente de manera irregular en la atmósfera (Harrison, 2006). Una medida de las actividades humanas está asociada al día de la semana, puesto que el tráfico vehicular aumenta durante los días laborales y también los niveles de concentración de los contaminantes atmosféricos (Ballester, Tenías y Pérez-Hoyos, 1999). Paralelamente, la temperatura, humedad, presión y altitud provocan condiciones atmosféricas inestables, con presencia de inversiones térmicas que hacen que las partículas a nivel del suelo se concentren por más tiempo y, por lo tanto, afecten la calidad del aire sobre las aglomeraciones urbanas. Por otro lado, factores meteorológicos como la lluvia arrastran las partículas contaminantes presentes en el aire, dependiendo de su duración, velocidad e intensidad y solubilidad, tamaño y velocidad de caída. Este fenómeno genera un lavado de la atmósfera, pero no indica la eliminación de partículas contaminantes, sino la transformación y el traslado a otros lugares como el suelo, vegetación y masa de agua en la superficie terrestre (Seoánez, 2002).

En las últimas décadas se han incrementado los problemas de contaminación atmosférica como consecuencia de diversas actividades antropogénicas. Por tal motivo, ha sido de gran interés estudiar los impactos ambientales de nivel local, regional y mundial en la salud humana. De acuerdo con los reportes presentados por el IDEAM (2007), esta situación no es ajena a las ciudades colombianas. En Cali, por ejemplo, el DAGMA (2012) ha reportado que las concentraciones de PM10 (24 horas), en promedio, se encontraron por debajo de la norma, sin embargo, se presentaron casos de PM10 (24 horas) que sobrepasaron los niveles permitidos por la Resolución 610 de 2010 del Ministerio de Ambiente, Vivienda y Desarrollo Territorial (concentraciones por encima de 100 $\mu g/m^3$).

Conocer el comportamiento horario del material particulado constituye una herramienta útil para la toma de decisiones en materia de calidad del aire. Esta información se obtiene por medio de una caracterización de la dinámica de las concentraciones de PM10 en el tiempo, para identificar las horas del día en las cuales se presentan graves episodios de contaminación. Algunos trabajos (Varó y Carratalá, 2002; Bedoya y Martínez, 2009) han mostrado resultados sobre el comportamiento horario de diversos contaminantes atmosféricos mediante curvas de valores promedio. Estos resultados no van más allá de un análisis meramente descriptivo, por carecer de un análisis estadístico más riguroso. Montoya, Morales y Olaya (2005) muestran que la regresión no paramétrica es una técnica robusta en relación con los estudios de calidad del aire, puesto que permite la modelación del comportamiento típico de un contaminante a lo largo de un día sin necesidad de hacer supuestos sobre la forma funcional de los datos.

El propósito de este trabajo es proponer un modelo no paramétrico para el comportamiento horario de las concentraciones de partículas PM10 en el aire. La propuesta se ilustra con datos de la zona norte de Cali, Colombia, teniendo en cuenta el día de la semana y la precipitación. Dado que se

tienen múltiples registros de las concentraciones de PM10 por hora, se realizaron comparaciones entre dos estimadores no paramétricos *kernel* y *spline*, los cuales permiten considerar el escenario de múltiples respuestas por punto de diseño.

2. DATOS

Dos conjuntos de datos fueron utilizados en este estudio. El primero corresponde solo a los registros horarios de las concentraciones promedio de PM10 vigilados por la Red de Monitoreo de la Calidad del Aire (RMCA) del DAGMA en la estación Éxito de La Flora. El segundo corresponde a los registros diarios de lluvia llevados por el IDEAM en la estación Sede IDEAM. Ambos corresponden al periodo de observación febrero-diciembre de 2010 en el norte de la ciudad de Cali.

Los niveles de concentración de PM10 son valores promediados cada hora, medidos en microgramos por metro cúbico ($\mu g/m^3$). El sistema de seguimiento toma 360 datos por hora, información recolectada cada 10 segundos. Por otro lado, los registros de precipitación corresponden al volumen de agua en milímetros (mm) que cae en un periodo por m^2. El sistema de seguimiento automático registra la cantidad de lluvia diaria acumulada cada 24 horas a las 7 a. m.

Para efectos de análisis, la hora en la cual se captó el registro de los valores de PM10 fue modificada a configuración numérica, de tal manera que la hora 0 representa la hora 12:00 a.m. (medianoche) y la hora 23, la última hora del día, 11:00 p. m.

2.1 Regresión no paramétrica

La regresión no paramétrica, al igual que la paramétrica, permite estimar el valor promedio de una variable respuesta en función de una o más variables predictoras (Härdle, 1992). Es importante destacar que, en muchos casos, esta relación no se comporta siempre de forma lineal, como ocurre con los contaminantes atmosféricos, cuyo comportamiento es complejo e influenciado por factores meteorológicos, características de las fuentes de emisión y aspectos topográficos (Harrison, 2006). En este sentido, la aplicación de un modelo paramétrico pierde firmeza en comparación con las técnicas de regresión no paramétrica, las cuales modelan el comportamiento de un conjunto de datos sin asumir a priori una forma funcional conocida (Bowman y Azzalini, 1997).

Diversos autores (Nadaraya, 1964; Watson, 1964; Priestley y Chao, 1972; Cleveland, 1979; Gasser y Müller, 1984) han propuesto estimadores comúnmente llamados *suavizadores* para estimar funciones de regresión no paramétrica. La forma y suavidad de una función estimada depende en gran medida del parámetro de suavización λ, el cual se escoge a partir de una medida que equilibre el sesgo y varianza de esta clase de estimadores (Härdle, 1992).

Al igual que en los modelos de regresión clásica, una forma general de expresar el modelo de regresión no paramétrica está dada por:

$$y=\mu(x)+\varepsilon \qquad (1)$$

donde y es la variable de respuesta aleatoria, $\mu(x)$ es una función suave desconocida, x es la variable independiente o de diseño y ε son los errores del modelo con media cero y varianza común σ^2. La regresión no paramétrica se ha empleado para encontrar asociaciones entre la contaminación del aire y enfermedades de tipo cardiovascular y respiratorio (Pope y Dockery, 2006), pero muy poco en la modelación del comportamiento horario de contaminantes atmosféricos, incluido el PM10. En este sentido, las técnicas de suavización *kernel* y *spline* ofrecen algunos estimadores que permiten modelar una variable de interés cuando se tienen varias observaciones por punto de diseño. En las siguientes secciones se mencionarán las ventajas y desventajas en la aplicación de estas técnicas en esta clase de estudios.

2.1.1 Suavizamiento kernel

Los suavizadores *kernel* más comunes se conocen como estimadores lineales respecto a los datos o respuestas y_i, los cuales tienen la siguiente forma:

$$\hat{\mu}(x) = n^{-1} \sum_{i=1}^{n} W(x, x_i, \lambda) y_i \quad (2)$$

donde

$$W(x, x_i, \lambda) = n^{-1} K\left(\frac{x - x_i}{\lambda}\right), \quad u = \frac{x - x_i}{\lambda} \quad (3)$$

La función $W(x, x_i, \lambda)$ pondera las observaciones de la variable respuesta paralela a los valores de la variable de diseño que están lejanas o cercanas a un punto de evaluación x. Esta colección de pesos hace uso de funciones tipo *kernel* $K(u)$ que son simétricas alrededor de cero y decrecen a medida que los valores de x_i se alejan del punto de evaluación x. Las funciones *kernel* más conocidas son la función gaussiana, triangular, uniforme, de Epanechnikov y biponderada (Härdle, 1992). La selección de la función *kernel* que se utilizará tiene muy poca importancia en la estimación de la función de regresión, siempre y cuando se garantice que el parámetro de suavización sea obtenido mediante la minimización del error cuadrático integrado medio (MISE) (Härdle, 1992).

Existen varios estimadores *kernel* de la función de regresión no paramétrica. Las diferencias que se pueden encontrar entre estos estimadores tienen que ver básicamente con propiedades relacionadas con el espaciamiento de los datos, la aleatoriedad de la variable predictora y la garantía de que la suma de los pesos asignados por la función *kernel* sea igual a uno.

El estimador propuesto por Nadaraya (1964) y Watson (1964) puede ser mirado como la idea básica del suavizamiento. Es necesario aclarar que este estimador es eficiente cuando la variable explicativa o de diseño es aleatoria. Años más tarde, este estimador fue extendido por Benedetti (1975) para el caso de diseños fijos.

Otra clase de estimadores *kernel* fue introducida por Cleveland (1979), quien propuso un *estimador de regresión localmente*. Su construcción se genera a partir de una solución mediante mínimos cuadrados ponderados en un polinomio de grado d:

$$min \sum_{i=1}^{n} W(x, x_i, \lambda)(y_i - \beta_0 - \beta_1(x_i - x) - \cdots - \beta_d(x_i - x)^d)^2 \quad (4)$$

Cuando el polinomio es de grado $d=1$, el estimador construido en (4) se conoce como el *estimador lineal local* (Bowman y Azzalini, 1997), que tiene ventajas sobre otros estimadores tipo *kernel*, ya que a medida que el parámetro de suavización aumenta, los pesos dados por la función *kernel* se acercan más y la curva estimada se aproxima a una línea de regresión por mínimos cuadrados (Cleveland, 1979). El estimador lineal local es mucho más eficiente en los bordes de la región de la covariable sobre la cual los datos fueron generados. Fan y Gijbels (1992) y Fan (1993) muestran excelentes propiedades de este estimador y concluyen que hay situaciones donde es preferible utilizar polinomios locales de grados impares.

Es importante tener en cuenta que los estimadores *kernel* mencionados fueron construidos con un modelo homocedástico y en el cual se tiene una sola observación y_i asociada a cada x_i. Sin embargo, intentar ajustar una curva de regresión a un conjunto de datos cuando se tienen múltiples respuestas por cada punto de diseño, a partir de los estimadores mencionados, no resulta adecuado cuando se tiene como objetivo realizar inferencias en el problema en estudio.

Una primera aproximación a este problema fue propuesta por Bowman y Azzalini (1997) quienes trabajaron con el escenario de medidas repetidas, donde cada individuo o sujeto es observado en diferentes puntos del tiempo. Para ajustar un modelo de regresión no paramétrica, estos autores toman en cuenta a los individuos como perfiles que presentan algún grado de correlación, pero que son independientes entre sí. En este sentido, Bowman y Azzalini (1997) contemplan el siguiente modelo de regresión:

$$y_{it} = \mu(x_i) + \varepsilon(x_{it}),\ i = 1,2,\ldots,N\ t = 1,2,\ldots,k \quad (5)$$

donde y_{it} es la variable repuesta, x_{it} es la variable explicativa o de diseño fija, $\mu(x_i)$ es la función media o curva de regresión y ε es el término error, cuya estructura de dependencia es de la forma:

$$cov(\varepsilon_{it}, \varepsilon_{kl}) = \begin{cases} \sigma^2 \rho(x_t - x_l) & si \quad i = k \\ 0 & si \quad i \neq k \end{cases} \tag{6}$$

siendo σ^2 la varianza del proceso y ρ_0=1. La estimación de $\mu(x)$ se obtiene a partir de los promedios de la variable respuesta por cada punto de diseño. Bowman y Azzalini (1997) muestran que el valor esperado de la función de regresión expresada en (5) no se ve afectado por la estructura de correlación, mientras que la varianza del estimador sí se afecta por la matriz de varianzas-covarianzas *V*. El sesgo y varianza del estimador están expresados de la siguiente forma:

$$E(\hat{\mu}) = S\mu, \quad Var(\hat{\mu}) = N^{-1}SVS^T \tag{7}$$

La matriz *S* denota la matriz de suavización, similar a la matriz sombrero (hat) utilizada en regresión paramétrica para la estimación de la función de regresión. La matriz de varianzas-covarianzas *V* se obtiene usando los residuales $\varepsilon_{it} = y_{it} - \bar{y}_t$. Si se desea construir intervalos de confianza o hacer pruebas de hipótesis, la estimación de la varianza de los errores $\hat{\sigma}^2$ no resulta ser la más conveniente, por lo que los residuales no dependen del parámetro de suavización (Eubank, 1999). Algunos criterios de estimación de la varianza residual se basan en diferencias sucesivas, pero no contemplan el escenario de medidas repetidas. Seifert, Gasser y Wolf (1993) presentan algunas sugerencias para la estimación de la varianza, las cuales se presentarán en la sección 2.1.3.

2.1.2 Suavizamiento spline

Eubank (1999) sugiere analizar los sujetos o individuos bajo estudio como múltiples respuestas y no como medidas repetidas o datos longitudinales. Las estimaciones llevadas a cabo en este caso se basan en el uso de *splines*. Eubank (1999) propone usar un *spline* cúbico (la versión más sencilla de los *splines*) en el caso de múltiples respuestas, donde la estimación de la función media se efectúa mediante un estimador de *f* que minimice la siguiente expresión:

$$N^{-1} \sum_{i=1}^{n} w_i \left(\bar{y}_i - f(x_i)\right)^2 + \lambda \int_0^1 f''(x)^2 \, dx \tag{8}$$

Nótese que la expresión del lado izquierdo de la suma en (8) hace referencia a una medida estándar de bondad de ajuste a los datos y la expresión del lado derecho representa una medida de la suavidad asociada a una función *f* que pertenece al espacio de funciones de Sobolev W_2^2 [0,1], cuyas segundas derivadas son de cuadrado integrable y donde [0,1] es un intervalo que contiene los puntos de diseño. Aquí, $\bar{y}_i$ corresponden a los promedios en cada punto de diseño y $w_i = n_i/s_i^2$, $i = 1, \ldots, n$ son pesos positivos que resultan adecuados para el caso de observaciones heterocedásticas.

El minimizador de *f* en (8) es en esencia un estimador de series de cosenos ponderados, equivalente asintóticamente a un estimador kernel. De esta manera sus propiedades de consistencia permiten seleccionar a λ mediante un estimador que minimiza el riesgo o pérdida. Cuando el parámetro de suavización λ tiende a infinito, las estimaciones tenderán a "sobresuavizarse" y cuando λ tiende a cero, se producirá un estimador que interpolará los datos. Una forma de seleccionar el parámetro de suavización es mediante el criterio de validación cruzada generalizada (CVG), expresado así:

$$CVG(\lambda) = \frac{N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} s_i^{-2} (y_{ij} - f(x_i))^2}{[1 - N^{-1} trS]^2} \tag{9}$$

Nótese que $CVG(\lambda)$ no necesita la estimación de la varianza de los errores para estimar el parámetro de suavización λ. La selección entre un método y otro radica más en problemas estadísticos que computacionales.

2.1.3 Inferencia en regresión no paramétrica

Para poder realizar inferencias en regresión no paramétrica, es necesario hablar sobre la estimación de la varianza de los errores $\hat{\sigma}^2$ para la construcción de intervalos de confianza y pruebas de hipótesis. De acuerdo con lo anterior, varios autores han propuesto estimadores a partir de diferencias sucesivas, debido al sesgo implícito en los estimadores no paramétricos de la función de regresión. Una

solución a este problema primero fue dada por Rice (1984), quien utilizó la idea de diferenciación de primer orden (r = 1). Dos años más tarde, Gasser, Sroka y Jennen-Steinmetz (1986) propusieron un estimador basado en la idea de diferenciación de segundo orden (r = 2), con el objetivo de remover efectos de tendencia local. Dette, Munk y Wagner (1998) sugirieron emplear el estimador propuesto por Gasser, Sroka y Jennen-Steinmetz (1986) en casos donde la función de regresión sea periódica, ya que los estimadores como el propuesto por Rice (1984) se ven influidos por fluctuaciones bruscas de la función de regresión, lo cual lleva a la inflación de la varianza.

Hall, Kay y Titterington (1990) sugirieron una generalización de los estimadores de varianza propuestos, basada en diferencias sucesivas óptimas asintóticamente. El estimador propuesto por estos autores se expresa de esta forma:

$$\hat{\sigma}^2_{HKT} = \frac{1}{(n-r)} \sum_{i=2}^{n-r} \left(\sum_{k=0}^{r} d_{ik} y_{i+k} \right)^2 \qquad (10)$$

donde los coeficientes d_{ik} se calculan matemáticamente (Hall, Kay y Titterington 1990, Apéndice 3) con las siguientes condiciones:

$$\sum_{k=1}^{r} d_{ik} = 0 \quad y \quad \sum_{k=1}^{r} d_{ik}^2 = 1 \qquad (11)$$

El estimador propuesto por Hall, Kay y Titterington (1990) resulta adecuado en los casos en que el diseño sea equidistante y se tengan tamaños grandes de muestras. Dette, Munk y Wagner (1998) advierten que en muchas situaciones es más apropiado estimar la varianza residual a partir de estimadores basados en diferencias ordinarias, debido a que el control del sesgo es mucho mejor y, por lo tanto, se tiene en general un buen rendimiento.

Los estimadores de varianza basados en diferencias sucesivas tienen la particularidad de realizarse bajo el supuesto de una respuesta y_i asociada a cada valor x_i. Al igual que la estimación de la función de regresión, es importante considerar el efecto de poseer múltiples respuestas por cada punto de diseño. Una aproximación a este problema se presenta en Seifert, Gasser y Wolf (1993), quienes sugieren una estimación de la varianza considerando el efecto de la varianza entre y dentro de los k puntos de diseño.

Seifert, Gasser y Wolf (1993) proponen construir un estimador mixto $\hat{\sigma}^2_{MIX} = as^2 + (1-a)\hat{\sigma}^2$, donde su error cuadrático medio sea minimizado bajo la siguiente expresión:

$$a = \frac{ECM(\hat{\sigma}^2)}{ECM(s^2) + ECM(\hat{\sigma}^2)} \qquad (12)$$

Nótese que el estimador s^2 controla el efecto de la varianza dentro de los puntos de diseño, como en el análisis de varianza convencional ANOVA, y el estimador de varianza no paramétrico $\hat{\sigma}^2$ construido a partir de las medias muestrales $\bar{y}_i$ controla el efecto de la varianza entre los puntos de diseño. Estos autores recomiendan construir los seudorresiduales como en (10) y modificar la condición de la derecha de (11) por la siguiente expresión:

$$\sum_{k=1}^{r} \frac{d_{ik}^2}{n_{i+k}} = 1 \qquad (13)$$

A partir de las consideraciones mencionadas, el estimador mixto fue construido teniendo en cuenta combinaciones lineales de los coeficientes d_{ik} para tamaños de muestra diferentes en cada punto de diseño cuando el orden de diferenciación es $r = 2$.

3. METODOLOGÍA ESTADÍSTICA

Las curvas de suavización de las concentraciones de PM10 se estimaron teniendo en cuenta como variable respuesta la concentración promedio horaria de PM10 y la variable de diseño como la hora del día (x=0,1,2,…,23). Paralelamente, se construyeron bandas de variabilidad asociadas a las estimaciones efectuadas, como indican Bowman y Azzalini (1997), las que hacen referencia a intervalos de confianza punto a punto para $E(\hat{\mu}(x))$ en vez de $\hat{\mu}(x)$ y que además ayudan a interpretar si hay diferencias entre dos curvas en puntos particulares.

Para la suavización de las curvas típicas del comportamiento horario de PM10 por el estimador propuesto por Eubank (1999) se utilizó la función *smooth.spline* de la librería *stats*, la cual ajusta un *spline* cúbico a los datos, teniendo en cuenta múltiples valores observados de las concentraciones y_i en cada hora del día x_i. Para la suavización de las curvas típicas por el estimador propuesto por Bowman y Azzalini (1997) se usó la función *sm.rm* de la librería *sm*, que estima los perfiles medios de una matriz que se supone que contiene mediciones repetidas a partir de un conjunto de individuos.

El parámetro de suavización para el ajuste del *spline* cúbico a los datos fue seleccionado mediante el criterio de validación cruzada generalizada (CVG). Para la construcción de las bandas de variabilidad fue necesario crear una función en R que estimara la varianza residual mediante una modificación del estimador mixto propuesto por Seifert, Gasser y Wolf (1993), teniendo en cuenta los tamaños de muestras diferentes en cada punto de diseño x_i y la estimación de los coeficientes d_{ik} a partir de las consideraciones dadas por los mismos autores.

4. RESULTADOS

Partiendo del análisis exploratorio de datos, se encontró que las concentraciones de PM10 poseen un comportamiento característico a lo largo de un día típico en el norte de Cali. La figura 1 ilustra el diagrama de cajas y bigotes de las concentraciones de PM10 por hora, en la estación Éxito, durante el año 2010. Se observa un comportamiento no lineal durante las 24 horas, muy posiblemente por el efecto causado por las fuentes vehiculares en este horario. Nótese además una clara indicación de heterocedasticidad de las concentraciones de PM10 cuyas varianzas tienden a cambiar en diferentes horas del día.

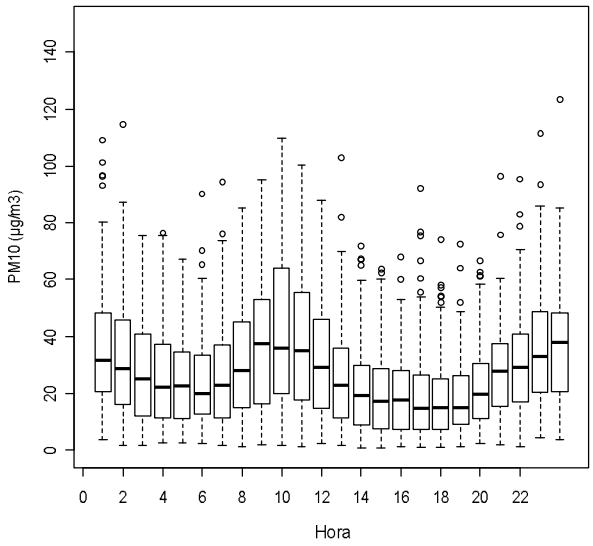

Figura 1. Diagrama de cajas y bigotes de las concentraciones promedio horarias de PM10 en la estación Éxito, año 2010

Se encontró también que las concentraciones del PM10 tienen un comportamiento condicionado por el día de la semana (ver figura 2). Esto permitió el agrupamiento de días, para posteriores análisis en la modelación horaria del PM10. En la figura 2, se observa que los días martes, miércoles, jueves, viernes y sábados poseen comportamientos similares, lo cual llevó a que se unieran en un solo grupo. Del mismo modo, los análisis para los días lunes y domingos y festivos fue realizado por separado.

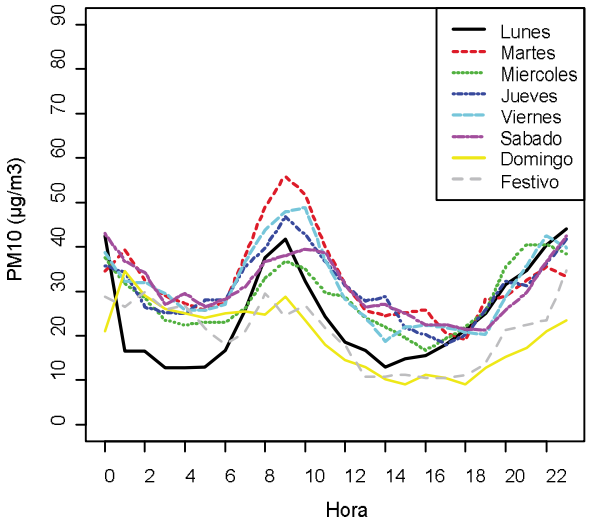

Figura 2. Concentraciones promedio horarias de PM10 en la estación Éxito, por días de la semana, año 2010

Para tener una aproximación del comportamiento del PM10 en temporadas secas y lluviosas, se procedió a analizar las concentraciones de material particulado por medio de la variable precipitación, donde se clasificó días sin lluvia, cuando la precipitación fuera igual a cero ($pp = 0\ mm$) y días con lluvia, como aquellos en los que la precipitación fuera mayor de cero ($pp > 0\ mm$). La figura 3 ilustra las concentraciones horarias de PM10 por temporada en la estación Éxito durante el año 2010, indicando aparentemente que los días en que ocurre lluvia, los niveles de contaminación de PM10 tienden a disminuir.

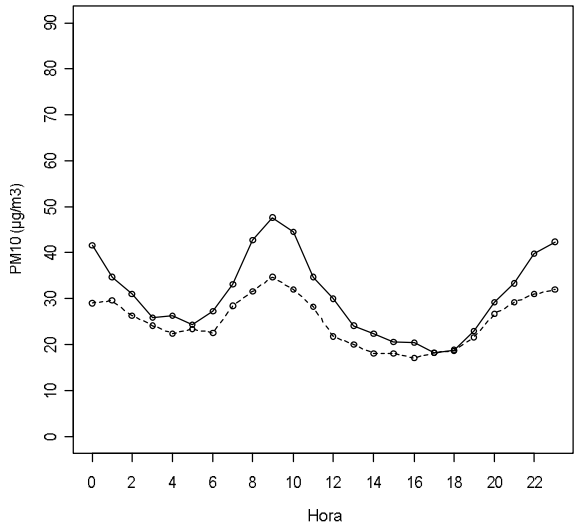

Figura 3. Concentraciones promedio horarias de PM10 en la estación Éxito, días sin lluvia (línea continua) y días con lluvia (línea segmentada), año 2010

Los resultados encontrados no dan información sobre cuál podría ser la distribución del comportamiento de las concentraciones de PM10, si se deseara ajustar un modelo de regresión paramétrica. No se pretende ajustar un modelo de pronóstico a las concentraciones de PM10, debido a que la regresión no paramétrica no involucra la estimación de parámetros que permiten predecir un valor de y dado un valor de x, sino que la función de regresión se construye a partir de lo que muestran los datos. Tampoco se implementó un análisis de series de tiempo, ya que esta técnica no permite modelar el comportamiento de una variable de interés que contemple múltiples respuestas por cada punto de diseño, ya que se requieren tantos modelos como horas-días por estimar y, a su vez, se presenta dificultad en la validación de supuestos, debido al gran conjunto de modelos paramétricos estimados (Barrientos, Olaya y González 2007).

La literatura indica que los métodos de suavización resultan ser los más adecuados cuando la distribución de una variable no sigue una tendencia lineal. Por lo tanto, se propone caracterizar el comportamiento horario de PM10 sin fines de pronóstico.

Teniendo en cuenta el escenario anterior, se propuso comparar los métodos propuestos por Bowman y Azzalini (1997) y Eubank (1999), para encontrar el modelo más adecuado para la estimación de las curvas típicas del PM10. Por lo tanto, se ajustaron curvas típicas horarias por tipo de día (lunes, martes-sábados y domingos-festivos) y temporada (días con lluvia y sin lluvia). Por otro lado, se construyeron bandas de variabilidad asociadas a las estimaciones realizadas, como indican Bowman y Azzalini (1997), las cuales ayudan a interpretar si hay diferencias entre dos curvas en puntos particulares.

Las estimaciones de las curvas de las concentraciones de PM10 por tipo de día se presentan en la figura 4, donde se observa que el comportamiento del PM10 resulta ser diferente al separar el análisis para días lunes, martes-sábados y domingos-festivos. Los días lunes se caracterizan por presentar una contaminación baja en las horas de la madrugada, en armonía con la reducción de las actividades laborales y con el éxodo de las personas en la víspera, mientras que los niveles de contaminación de PM10 son mayores en los días martes a sábados en relación con el aumento de las actividades laborales y el transporte vehicular, y los días domingos y festivos presentan concentraciones de PM10 bajas a partir de las 10 de la mañana, muy posiblemente por el cese de actividades laborales y el bajo flujo vehicular.

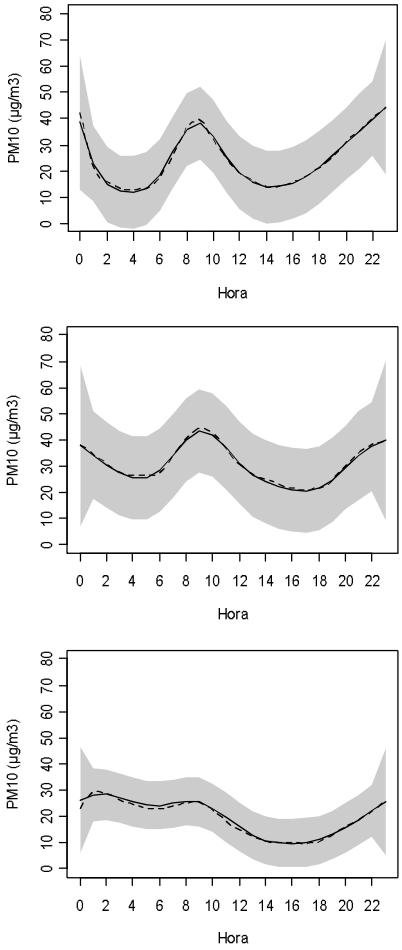

Figura 4. Curvas típicas de las concentraciones de PM10 y bandas de variabilidad para los días lunes (arriba), días martes-sábados (centro) y días domingos-festivos (abajo) en la estación Éxito, año 2010, usando el estimador *kernel* (línea segmentada) y *spline* cúbico (línea continua)

El aumento en la concentración de PM10 durante las horas de la noche en los tres tipos de días puede estar relacionado con la inversión térmica que ocurre durante esta franja horaria. Cuando las fuentes industriales y vehiculares emiten material particulado a la atmósfera en condiciones de inversión térmica, los niveles de concentración se elevan y generan episodios de contaminación que ponen en riesgo la salud humana.

Usando los modelos y bandas de variabilidad asociadas a las curvas de PM10 por tipo de día, la concentración de PM10 a las 10 a. m. en la estación Éxito de La Flora, se estima en $38{,}3 \pm 13{,}7\ \mu g/m^3$ para los días lunes, $43{,}3 \pm 15{,}9\ \mu g/m^3$ para los días martes a sábado y $25{,}4 \pm 9{,}3\ \mu g/m^3$ para los domingos y festivos, con un nivel de confianza del 95 %.

Otro aspecto importante en la modelación de las concentraciones horarias de material particulado es el efecto generado por la precipitación. El tamaño del material particulado PM10 hace que su proceso de depositación húmeda sea muy bajo, permaneciendo por largos periodos suspendido en el aire (Baird, 2001). Las estimaciones de las curvas de las concentraciones de PM10 por días con lluvia y sin lluvia son presentadas en la figura 5. Se observa aparentemente que las concentraciones de PM10 son un poco más altas en los días sin lluvia.

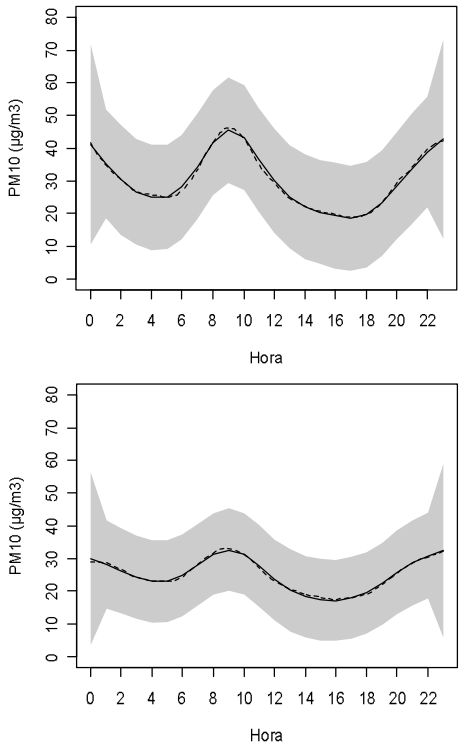

Figura 5. Curvas típicas de las concentraciones de PM10 y bandas de variabilidad para los días sin lluvia (arriba) y días con lluvia (abajo) en la estación Éxito, año 2010, usando el estimador *kernel* (línea segmentada) y *spline* cúbico (línea continua)

Cuando ocurre lavado de la atmósfera, las partículas contaminantes del aire precipitan y ayudan a minimizar el material particulado proveniente de fuentes industriales y vehiculares, lo cual a su vez reduce riesgos en la salud humana. Pero si los niveles de concentración de PM10 son altos, se altera la acidez de la lluvia causando graves consecuencias al ambiente (agua, suelo, vegetación y edificios) (Baird, 2001).

Usando los modelos y bandas de variabilidad de PM10 por días con lluvia y sin lluvia, la concentración a las 10 a. m. en la estación Éxito de La Flora en el 2010 se estima en $45{,}5 \pm 16{,}14\ \mu g/m^3$ para días con lluvia y $32{,}7 \pm 12{,}5\ \mu g/m^3$ para días sin lluvia, con un nivel de confianza del 95 %. Las bandas de variabilidad parecen indicar que no hay diferencias significativas del comportamiento del PM10 durante las temporadas secas y lluviosas en esta zona de observación, por lo que las bandas de la curva estimada para los días sin lluvia cubren la totalidad de las bandas de la curva estimada para los días con lluvia.

Nótese que en la figura 4 y figura 5 el estimador *kernel* basado en la idea de los perfiles propuesta por Bowman y Azzalini (1997) no apoya el supuesto de homocedasticidad de las observaciones, en comparación con el estimador *spline* cúbico propuesto por Eubank (1999), el cual genera pesos más pequeños al ajuste realizado en casos donde se presentan varianzas muestrales grandes.

5. DISCUSIÓN

La variación cíclica del PM10 en el transcurso de las 24 horas del día y los picos de contaminación en horas de la mañana reafirman la hipótesis de que las fuentes vehiculares son el principal responsable de los altos niveles de contaminación por PM10 en la zona norte de Cali. Por tanto, es importante tener en cuenta el comportamiento de la dinámica del tráfico en esta zona de seguimiento, para conocer cómo los episodios de contaminación varían a lo largo de un día de acuerdo con los cambios del flujo vehicular.

La contaminación por material particulado en la zona norte de Cali parece mostrar que las concentraciones de PM10 no varían por días secos y lluviosos. Un estudio posterior podría ser encaminado a contrastar esta hipótesis y evaluar el efecto de otros factores meteorológicos, en especial velocidad y dirección del viento, en la contaminación por material particulado mediante técnicas de suavización multivariadas que permitan explicar de forma más completa el comportamiento del PM10 en la ciudad de Cali.

Los resultados muestran indicios de que el día de la semana es un factor clave en la modelación horaria del material particulado. Por ello, es importante ajustar modelos de regresión teniendo en cuenta que las concentraciones de PM10 en los días lunes resulta ser diferente de los demás días ordinarios, con el fin de evitar factores de confusión en estudios sobre la calidad del aire por material particulado. Otra opción es introducir el tipo de día como una variable dentro de un modelo de regresión múltiple no paramétrica, por ejemplo, un modelo aditivo generalizado (Hastie y Tibshirani, 1990).

Las técnicas de suavización no paramétrica permiten avanzar en la construcción de modelos que caractericen el comportamiento diario de los contaminantes atmosféricos, en un escenario de múltiples respuestas por cada punto de diseño. Los resultados revelan que las estimaciones de PM10 más robustas se consiguen a partir de los suavizadores *spline*, que resultan más adecuadas cuando se dispone de más de una observación de PM10 por hora y que las estimaciones no se vean afectadas por la presencia de heterocedasticidad en las observaciones.

Las dificultades de los modelos *spline* utilizados en este trabajo subyacen en ignorar la correlación serial que puedan presentar las concentraciones de PM10 tomadas a través del tiempo. Existen algunos avances en la estimación de la estructura de dependencia de los errores en la construcción de modelos de regresión no paramétrica. Una de ellas es la expuesta por Bowman y Azzalini (1997), que

proponen estimar la función de regresión mediante suavización *kernel* suponiendo que la estructura de correlación de los errores sigue un proceso AR(1). Limitar las estimaciones con esta consideración no resulta adecuado, ya que pueden presentarse casos donde los errores sigan un proceso autorregresivo de diferente orden. Este es un problema abierto.

Otras soluciones se han ido dirigiendo al problema de la selección del parámetro de suavización cuando hay correlación serial de los errores, pero muy pocas a la especificación de una estructura de dependencia de los errores cuando se tienen múltiples respuestas por cada punto de diseño. Trabajos como los de Wang (1998) y Opsomer, Wang y Yang (2001) proponen estimar la función de regresión a partir de la selección óptima del parámetro de suavización cuando los errores están correlacionados. Sin embargo, estas propuestas no tienen en cuenta la especificación de una estructura residual que permita la estimación de las funciones de covarianzas y autocorrelación a los datos, siendo un problema que debe ser explorado con más detenimiento.

6. CONCLUSIONES

La principal conclusión se centra en el hecho de que los modelos *spline* resultan ser adecuados para modelar el comportamiento diario del material particulado. De esta forma podría aplicarse a cualquier contaminante atmosférico y ser una herramienta estadística que permita la estimación de concentraciones en cualquier hora del día, incluyendo aquellas en las que no se toman mediciones o, como en este caso, ajustar una curva para estudiar el comportamiento diario de un contaminante. Un uso potencial de este tipo de ajustes sería la construcción de datos funcionales que sirvan como datos diarios para alimentar un modelo de regresión funcional con una respuesta funcional, que permita estudiar el comportamiento diario, y variables de predicción funcionales.

Se estimaron en forma satisfactoria las curvas suaves de PM10 para días lunes, martes-sábados y domingos-festivos y días con lluvia y sin lluvia, siendo los días martes-sábados y los días sin lluvia aquellos en los cuales se presentó una mayor contaminación de material particulado en el norte de Cali.

El análisis de las bandas de variabilidad permitió identificar que en la zona norte de Cali es muy posible que no se presenten diferencias estadísticamente significativas de la contaminación por material particulado entre días secos y lluviosos. Esta observación requeriría una prueba formal, que no es uno de los objetivos de este trabajo.

AGRADECIMIENTOS

Los autores expresan agradecimientos a la Vicerrectoría de Investigaciones de la Universidad del Valle y al Programa Jóvenes Investigadores e Innovadores "Virginia Gutiérrez de Pineda" de Colciencias, por el apoyo financiero que hizo posible el desarrollo de este estudio. Igualmente agradecen al DAGMA y al IDEAM por proveer los datos utilizados en este proyecto, así como a los pares evaluadores cuyos valiosos aportes han mejorado la versión original.

REFERENCIAS

Baird, Colin. *Química ambiental*. Barcelona: Reverté, 2001. 622 p.

Ballester, Ferran; Tenías, José María y Pérez-Hoyos, Santiago (1999). "Efectos de la contaminación atmosférica sobre la salud: Una introducción". *Revista Española de Salud Pública*, vol. 73, No. 2 (marzo-abril), pp. 109-121.

Barrientos, Andrés Felipe; Olaya, Javier y González, Víctor Manuel (2007). "Un modelo spline para el pronóstico de la demanda de energía eléctrica". *Revista Colombiana de Estadística*, vol. 30, No. 2 (julio-diciembre), pp. 187-202.

Bedoya, Julián y Martínez, Elkin (2009). "Calidad del aire en el valle de Aburrá, Antioquia, Colombia". *Dyna*, vol. 72, No. 158 (mayo-agosto), pp. 7-15.

Benedetti, Jacqueline K. (1975). Kernel estimation of regression functions. Proceedings of the Computer Science and Statistics: 8th Annual Symposium on the Interface. Los Angeles (13-14 February), pp. 405-412.

Bowman, Adrian W. and Azzalini, Adelchi. *Applied smoothing techniques for data analysis: The kernel approach with S-Plus illustrations*. Oxford Statistical Science Series, 18. Oxford: Clarendon Press, 1997. 193 p.

Cleveland, William S. (1979). "Robust locally weighted regression and smoothing scatterplots". *Journal of the American Statistical Association,* vol. 74, No. 368 (December), pp. 829- 836.

Departamento Administrativo de Gestión Medio Ambiente (Cali) –DAGMA– (2012). *Boletín de la calidad del aire mes mayo de 2012*.

Dette, Holger; Munk, Axel and Wagner, Thorsten (1998). "Estimating the variance in nonparametric regression - What is a reasonable choice?". *Journal of the Royal Statistical Society*. Series B (Statistical Methodology), vol. 60, No. 4, pp. 751-764.

Eubank, Randall L. *Nonparametric regression and spline smoothing*. Statistics, Textbooks and Monographs. New York: Marcel Dekker, 1999. 338 p.

Fan, Jianqing (1993). "Local linear regression smoothers and their minimax efficiencies". *The Annals of Statistics*, vol. 21, No. 1 (March), pp. 196-216.

Fan, Jianqing and Gijbels, Irene (1992). "Variable bandwidth and local linear regression smoothers". *The Annals of Statistics*, vol. 20, No. 4 (December), pp. 2008-2036.

Gasser, Theo and Müller, Hans-Georg (1984). "Estimating regression functions and their derivatives by the kernel method". *Scandinavian Journal of Statistics*, vol. 11, No. 3, pp. 171-185.

Gasser, Theo; Sroka, Lothar and Jennen-Steinmetz, Christine (1986). "Residual variance and residual pattern in nonlinear regression". *Biometrika*, vol. 73, No. 3 (December), pp. 625-633.

Hall, Peter; Kay, J. W. and Titterington, D. M. (1990). "Asymptotically optimal difference-based estimation of variance in nonparametric regression". *Biometrika*, vol. 77, No. 3 (September), pp. 521-528.

Härdle, Wolfgang (1992). *Applied nonparametric regression*. Econometric Society Monographs. Cambridge, UK: Cambridge University Press, 333 p.

Harrison, Roy M. *An introduction to pollution science*. London: Royal Society of Chemistry, 2006. 322 p.

Hastie, Trevor J. and Tibshirani, Robert J. (1990). "Generalized additive models". Boca Raton, FL: CRC, 352 p.

Instituto de Hidrología, Meteorología y Estudios Ambientales –IDEAM– (2007). Informe nacional sobre calidad del aire, Colombia. .

Montoya, Martha Rocío; Morales, Alexandra y Olaya, Javier (2005). "Estimación no-paramétrica de curvas típicas diarias para los contaminantes CO, NO_2 y SO_2 en Santiago de Cali". *Revista Ingeniería de Recursos Naturales y del Ambiente,* vol. 2, No. 1, pp. 23-27.

Nadaraya, Elizbar A. (1964). "Some new estimates for distribution functions". *Theory of Probability and its Applications*, vol. 9, No. 3, pp. 497-500.

Opsomer, Jean; Wang, Yuedong and Yang, Yuhong (2001). "Nonparametric regression with correlated errors". *Statistical Science*, vol. 16, No. 2, pp. 134-153.

Perez-Padilla, R.; Schilmann, A and Riojas-Rodriguez, H. (2010). "Respiratory health effects of indoor air pollution". *The International Journal of Tuberculosis and Lung Diseases,* vol. 14, No. 9 (September), pp.1079-1086.

Pope, C. Arden and Dockery, Douglas W. (2006). "Health effects of fine particulate air pollution: Lines that connect". *Journal of the Air & Waste Management Association,* vol. 56 (June), pp. 709-742.

Priestley, M. B. and Chao, M. T. (1972). "Non-parametric function fitting". *Journal of the Royal Statistical Society*, Series B (Methodological) vol. 34, No. 3, pp. 385-392.

Rice, John (1984). "Bandwidth choice for nonparametric regression". *The Annals of Statistics*, vol. 12, No. 4 (December), pp. 1215-1230.

Seifert, Burkhardt; Gasser, Theo and Wolf, Andreas (1993). "Nonparametric estimation of residual variance revisited". *Biometrika*, vol. 80, No. 2, pp. 373-383.

Seoánez, Calvo, Mariano. *Tratado de la contaminación atmosférica: Problemas, tratamiento y gestión*. Colección Ingeniería del Medio Ambiente. Madrid: Mundi-Prensa, 2002. 1111 p.

Varó, Pedro y Carratalá, A. (2002). "Evolución de los niveles de inmisión de contaminación atmosférica en una ciudad industrial (Alcoy) desde 1989 a 2000". *Revista de Salud Ambiental,* vol. 2, No. 1, pp. 8-15.

Wang, Yuedong (1998). "Smoothing spline models with correlated random errors". *Journal of the American Statistical Association*, vol. 93, No. 441, pp. 341-348.

Watson, Geoffrey S. (1964). "Smooth regression analysis". *Sankhyā: The Indian Journal of Statistics*, Series A (1961-2002), vol. 26, No. 4 (December), pp. 359–372.