DYNA
ISSN: 0012-7353
Universidad Nacional de Colombia

Olaya-Ochoa, Javier; Ovalle, Diana Paola; Urbano, Cristhian Leonardo
Acerca de la estimación de la fracción PM 2.5 /PM 10
DYNA, vol. 84, núm. 203, 2017, Octubre-Diciembre, pp. 343-348
Universidad Nacional de Colombia

DOI: https://doi.org/10.15446/dyna.v84n203.65228

Disponible en: https://www.redalyc.org/articulo.oa?id=49655603044

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA
http://dyna.medellin.unal.edu.co/

On the $PM_{2.5}/PM_{10}$ fraction estimation

Javier Olaya-Ochoa, Diana Paola Ovalle & Cristhian Leonardo Urbano

Escuela de Estadística, Facultad de Ingeniería, Universidad del Valle, Cali, Colombia. javier.olaya@correounivalle.edu.co, diana.ovalle@correounivalle.edu.co, cristhian.urbano@correounivalle.edu.co

Received: May 24th, 2017. Received in revised form: October 30th, 2017. Accepted: November 3rd, 2017.

Abstract
We discuss the estimation of the $PM_{2.5}/PM_{10}$ fraction, which some authors report as the $PM_{2.5}/PM_{10}$ ratio. Previous studies highlight the importance of this fraction in the study of the impact of airborne particles on human health and their potential use in reconstruction and imputation of $PM_{2.5}$ fine particle hourly levels from the PM_{10} coarse particle hourly levels. We suggest adapting the estimation process using different strategies for particular situations. We used the general linear model (GLM), the regression through the origin model (RTO), the generalized additive models (GAM) and the regression models based on weighted least squares (WLS).

Keywords: Environmental pollution, fine particles, thick particles, GLM, RTO, GAM, WLS.

Acerca de la estimación de la fracción $PM_{2.5}/PM_{10}$

Resumen
Se discute la estimación de la fracción $PM_{2.5}/PM_{10}$, que algunos autores reportan como la razón $PM_{2.5}/PM_{10}$. Estudios previos destacan la importancia de esta fracción en el estudio del impacto de las partículas en el aire sobre la salud de las personas y su uso potencial en la reconstrucción y en la imputación de datos de los niveles de partículas finas $PM_{2.5}$ a partir de los niveles de partículas gruesas PM_{10}. Se sugiere adaptar la estimación, mediante el uso de diferentes estrategias para situaciones particulares, incluyendo el modelo lineal general (GLM), la regresión a través del origen (RTO), los modelos aditivos generalizados (GAM) y los modelos de regresión usando mínimos cuadrados ponderados (WLS).

Palabras clave: Contaminación ambiental, partículas finas, partículas gruesas, GLM, RTO, GAM, WLS.

1. Introducción

La presencia de contaminantes en el aire se encuentra relacionada con diferentes problemas de morbilidad y mortalidad, según lo ha definido la Organización Mundial de la Salud [13]. Dentro de esos contaminantes es de especial interés el material particulado (PM), el cual se clasifica en tres tipos: partículas gruesas (PM_{10}), finas ($PM_{2.5}$) y ultrafinas (PM_1). Las partículas gruesas son aquellas con un diámetro aerodinámico entre 2.5 y 10 µm; las finas tienen diámetro menor que 2.5 µm; y las ultrafinas menor que 1 µm [7].

En las definiciones anteriores es importante observar que el $PM_{2.5}$ no es una parte del PM_{10}, por lo que formalmente el cociente $PM_{2.5}/PM_{10}$ es una razón. Sin embargo, los equipos de lectura están diseñados para medir la cantidad de partículas en el aire que tienen un diámetro aerodinámico menor que 10 µm, por lo que en estas mediciones de partículas gruesas se incluyen las partículas finas de un diámetro aerodinámico menor que 2.5 µm. Es por esta razón que en este trabajo se prefiere hablar de la "fracción $PM_{2.5}/PM_{10}$", en lugar de hablar de la "razón $PM_{2.5}/PM_{10}$".

De acuerdo con Munir en 2017, el tamaño de las partículas en el aire es un factor decisivo para determinar el tiempo que permanecen las partículas en la atmósfera, así como para determinar en qué partes del tracto respiratorio se podrían depositar [11]. En este contexto, según Munir en 2017, la fracción $PM_{2.5}/PM_{10}$ es un indicador de la presencia de partículas finas en el aire que contribuye al estudio del tamaño de las partículas [11]. Gomiscek, en el 2004, se apoyan en estudios epidemiológicos generales para afirmar que la fracción de partículas finas en el aire tiene un impacto considerable sobre la salud humana, incluso en concentraciones por debajo de los límites establecidos para $PM_{2.5}$ en las normas internacionales [9]. Por otra parte, según la OMS en 2016, las partículas $PM_{2.5}$ se encuentran asociadas a la aparición de enfermedades respiratorias crónicas como

How to cite: Olaya-Ochoa, J., Ovalle, D.P. and Urbano, C.L., Acerca de la estimación de la fracción PM2.5/PM10 DYNA, 84(203), pp. 343-348, December, 2017.

cáncer de pulmón, asma y pulmonía, entre otras [13]. Debido a ello, se han implementado normas para la medición del contaminante, lo cual requiere protocolos de monitoreo y vigilancia permanente. Sin embargo, al contrario de lo que sucede con la medición de partículas PM_{10}, la medición de partículas $PM_{2.5}$ suele estar restringida en los sistemas de vigilancia de la calidad del aire en ciudades de países en vía de desarrollo, debido esencialmente a problemas de costos o de implementación técnica, por lo que la fracción $PM_{2.5}/PM_{10}$ sería útil además como herramienta de estimación de los niveles de $PM_{2.5}$ a partir de los niveles de PM_{10}. Esta herramienta podría utilizarse bien para imputar datos faltantes en una estación, o bien para efectos de estimación en estaciones de vigilancia en los que se mide PM_{10} pero no $PM_{2.5}$. En trabajos previos [4, 17], usan modelos de regresión lineal para estimar la fracción $PM_{2.5}/PM_{10}$ en las ciudades de Medellín y Bogotá (Colombia), respectivamente. Adicionalmente, el departamento de protección ambiental de Hong Kong [6] propone un método de estimación de la fracción $PM_{2.5}/PM_{10}$, citando a Smyth en el 2006 [19], pero no dan indicaciones de las propiedades del estimador propuesto.

Este artículo evalúa varias aproximaciones para la estimación de la fracción $PM_{2.5}/PM_{10}$, validando sus potencialidades como herramienta de estimación, iniciando desde un modelo lineal general (GLM) [8,1], un modelo de regresión a través del origen (RTO) [12] y un modelo aditivo generalizado (GAM) [10, 20]. En lo que sigue, se describen brevemente los métodos usados, se muestran los resultados de su aplicación a un conjunto de datos particular y se formulan algunas recomendaciones para posibles usos de estos modelos en este marco.

2. Materiales y métodos

Los datos que se utilizan para ilustrar este artículo fueron colectados en la ciudad de Santiago de Cali, Colombia. La ciudad cuenta con un Sistema de Vigilancia de la Calidad del Aire (SVCASC) a cargo del Departamento Administrativo de Gestión del Medio Ambiente [3] y cuenta con ocho estaciones fijas en operación, así como una unidad móvil, ubicadas en diferentes puntos de la ciudad. El sistema mide las partículas con diámetro menor a 10 μm incluyendo las partículas gruesas y finas y reporta este resultado como PM_{10}. La medición de las partículas finas se reportan en concordancia con la definición oficial de la EPA (2015) [7]. Cuatro de las nueve estaciones de vigilancia del SVCASC se encuentran habilitadas para la medición de partículas finas y en dos de estas se miden ambos contaminantes PM_{10} y $PM_{2.5}$; en este trabajo se usan los datos de ambos contaminantes en la estación Compartir.

La estación Compartir del SVCASC es la estación de vigilancia de la calidad del aire ubicada más al oriente en la ciudad de Santiago de Cali. Aún más al oriente de la estación, se encuentra una extensa planicie que va desde el Río Cauca hasta la cordillera central colombiana, planicie en la que predominan cultivos de caña de azúcar. No hay plantas industriales mayores en la zona y la generación de partículas suspendidas en el aire se puede asociar principalmente a las fuentes móviles y a emisiones de polvo de vías pavimentadas y sin pavimentar, a construcciones en la zona y a la erosión debida al viento en áreas cultivadas y sin cultivar [15]. En esta estación se cuenta con un total de 3,125 datos disponibles para el período Enero-Noviembre 2015. Esta es una cantidad mucho mayor a los datos con los cuales se realizaron los trabajos de Rojas en el 2005 [17] y Echeverri en el 2008 [4] y similar a los que se usaron en EPD en 2017 [6], aunque estos últimos se refieren a varios años. Los datos de otras estaciones se incluyen en algunos análisis adicionales. Todo el procesamiento de datos se hace usando el lenguaje y ambiente para computación estadística R [16].

2.1. Regresión lineal (GLM)

De acuerdo con Fox en el 2016, el análisis de regresión lineal simple examina la relación entre una respuesta $\boldsymbol{Y}$ y una variable de predicción $\boldsymbol{X}$ [8]. El modelo busca estimar la media de la respuesta $\boldsymbol{Y}$ para un valor dado de $\boldsymbol{X}$ y presume que esta media se encuentra sobre la recta de regresión. Según Agresti en el 2015, en el modelo lineal ordinario se asume que el vector $\boldsymbol{y} = (\boldsymbol{y_1}, \dots, \boldsymbol{y_n})$ tiene media $\boldsymbol{\mu} = (\boldsymbol{\mu_1}, \dots, \boldsymbol{\mu_n})$ y varianza $\boldsymbol{V} = \boldsymbol{\sigma^2 I}$, donde $\boldsymbol{I}$ es la matriz identidad de tamaño $\boldsymbol{n} \times \boldsymbol{n}$. Así que se busca estimar el vector de medias $\boldsymbol{\mu}$ que se representa como $\boldsymbol{\mu} = \boldsymbol{X\beta}$, con $\boldsymbol{X} = \boldsymbol{x_{ij}}$ la matriz de diseño y $\boldsymbol{\beta}$ un vector de parámetros de tamaño $\boldsymbol{p} \times \boldsymbol{1}$ con $\boldsymbol{p} \leq \boldsymbol{n}$ [1].

Una forma alternativa de expresar este modelo es:

$$Y = X\beta + \varepsilon \tag{1}$$

En esta versión del modelo, se presume que $\boldsymbol{\varepsilon}$ es un término de error tal que tiene $\boldsymbol{E(\varepsilon) = 0}$ y matriz de covarianza $\boldsymbol{var} = (\boldsymbol{\varepsilon}) = \boldsymbol{\sigma^2 I}$. [1].

Si se adiciona el supuesto de Normalidad de *Y* en la primera formulación del modelo o de $\boldsymbol{\varepsilon}$ en la segunda, entonces el modelo sería un Modelo Lineal Generalizado (GLM), que usa la identidad como función de enlace.

Bajo estas condiciones, la estimación de $\boldsymbol{\beta}$ se conduce mediante Mínimos Cuadrados Ordinarios (OLS). Pero si la matriz $\boldsymbol{V \neq \sigma^2}$, entonces β se deberá estimar usando Mínimos Cuadrados Ponderados (WLS). En el primer caso el estimador será $\boldsymbol{\widehat{\beta} = (X'X)^{-1}X'Y}$; en el segundo será $\boldsymbol{\widehat{\beta} = (X'VX)^{-1}X'VY}$

Para evaluar los supuestos de estos modelos se acude habitualmente a los residuales, definidos como $\boldsymbol{e = (I - H)Y}$, donde *H* se define como $\boldsymbol{H_{OLS} = X(X'X)^{-1}X'}$ en el modelo OLS y como $\boldsymbol{H_{wLS} = X(X'VX)^{-1}X'V}$ en el modelo WLS.

2.2. Regresión a través del origen (RTO)

En ocasiones se ajustan modelos que pasan por el origen (RTO), es decir, que no ajustan un intercepto. Estos modelos tienen una larga tradición de estudio en los círculos académicos. En 1983, el profesor George Casella publicó en The American Statistician un artículo [2], en el cual busca precisar en particular los mecanismos de comparación de modelos con y sin intercepto. 20 años después, Joseph Eisenhauer compara cómo se hace RTO en tres software de

uso común [5]. Luego en 2014, Othman retoma la discusión de Casella [14]. La discusión sigue abierta.

Este modelo puede expresarse así:

$$y_i = \beta x_i + \varepsilon_i \tag{2}$$

Para efectos computacionales, en esta versión del modelo se modifica la matriz de diseño *X* suprimiendo la primera columna. Pero este cambio modifica por completo el modelo, de tal manera que no resulta ser una tarea sencilla comparar este nuevo modelo con un modelo en el que se ha ajustado el intercepto. El coeficiente de determinación $\boldsymbol{R^2}$, para dar solo un ejemplo, cambia radicalmente y la forma de calcularlo para el modelo RTO no es comparable con el obtenido para el GLM. Los autores Eisenhauer, Casella y Othman [5, 2, 14] muestran que ajustar un modelo sin intercepto es equivalente a añadir un punto adicional al problema. Luego se verifica si este punto adicional tiene un leverage muy alto y, en tal caso, se concluye que el intercepto no se puede suprimir. El punto que se añade se denota $(\boldsymbol{n^*\bar{x}, n^*\bar{y}})$, con $\boldsymbol{n^* = n/(\sqrt{n+1}-1)}$. Otros autores [18] sugieren comparar estos dos modelos usando el criterio de información de Akaike (AIC).

2.3. Modelos aditivos generalizados (GAM)

Los GAM fueron propuestos a finales de los años 1990's [9] y han sido gradualmente actualizados por versiones posteriores [20]. En los GAM se reemplaza la combinación lineal $\boldsymbol{X\beta}$ de los GLM, que luciría en notación extendida como $\sum_{j=1}^{p} \boldsymbol{\beta_j x_j}$, por la suma de funciones suaves definida en la ecuación 3 donde $\boldsymbol{s_j}$ es una función obtenida mediante métodos de suavización.

$$\sum_{j=1}^{p} \boldsymbol{s_j}(\boldsymbol{x_j}) \tag{3}$$

Las funciones suaves s_j se obtienen como combinaciones lineales de un conjunto de funciones que forman una base de un espacio funcional. La matriz *H* se construye en este caso usando las realizaciones de las funciones de la base en los valores de *X*, en lugar de usar los valores de *X*.

3. Resultados

3.1. Los datos

El diagrama de dispersión de la Fig. 1 muestra la relación entre la concentración de PM_{10} y $PM_{2.5}$. El coeficiente de correlación lineal es 0.83. En el proceso de depuración se suprimieron las observaciones que mostraban niveles de $PM_{2.5}$ mayores que los niveles de PM_{10}. Los mínimos de ambas variables están muy cerca al origen, como podría anticiparse ya que si el $PM_{2.5}$ es una fracción del PM_{10}, entonces el $PM_{2.5}$ debería ser nulo si no se observara presencia en el aire de partículas PM_{10}.

3.2. Dos modelos lineales generales

El primer modelo lineal simple construido con los 3,124 datos disponibles en la estación Compartir del SVCASC

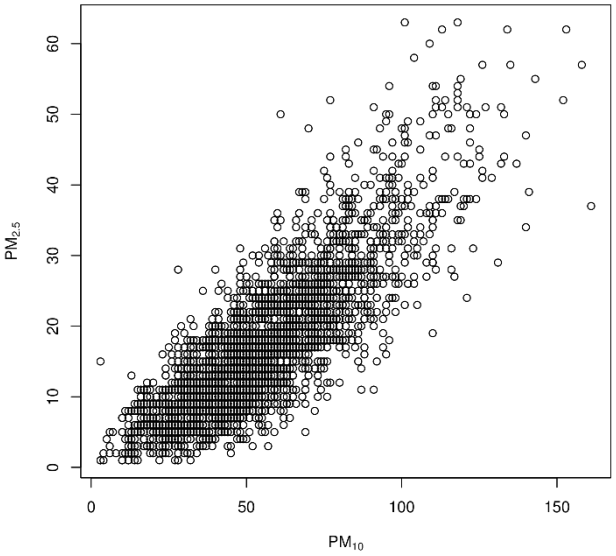

Figura 1. $PM_{2.5}$ vs PM_{10}.
Fuente: Elaboración Propia.

indica que los niveles de $PM_{2.5}$ (variable de respuesta *Y*) podrían estimarse a partir de los niveles horarios de PM_{10} (variable de predicción *X*), usando la ecuación $\boldsymbol{\hat{y}_i = -2.5975 + 0.3612 x_i}$

Los resultados de análisis indican que este modelo está en capacidad de explicar el 69.36% de la variación de las concentraciones de PM_{10} y que ambos estimadores son significativamente diferentes de 0. Si se asumiera la pendiente de este modelo como una aproximación a la fracción $PM_{2.5}/PM_{10}$ en esta estación, esto significaría que los niveles horarios de $PM_{2.5}$ se podrían estimar a partir de los de PM_{10} usando el multiplicador 0.3612. Pero esto implicaría ignorar el hecho de que el intercepto de la recta indica que para estimar un contaminante a partir del otro se debe restar la cantidad 2.5975 al producto del factor 0.3612 por la concentración de PM_{10}. De hecho este resultado conduce a una estimación negativa de la respuesta para valores de la variable de predicción incluso superiores a 7 µg/m^3.

Una opción sería construir un modelo sin intercepto. Como se observa en la Fig. 1, la nube de puntos se encuentra muy cercana al origen, por lo que no hay razón alguna para creer que el modelo dejará de ser lineal entre la nube de puntos y el origen. Por otra parte, siendo $PM_{2.5}$ una parte de PM_{10}, entonces si $\boldsymbol{PM_{10} = 0}$, se sigue que necesariamente $\boldsymbol{PM_{2.5} = 0}$. Y, por último, el parámetro β de un modelo del tipo 2 es evidentemente un estimador de la fracción $PM_{2.5}/PM_{10}$, evitando de paso las dificultades de estimación del modelo 1 ajustado a estos datos, que estima valores negativos de la concentración de $PM_{2.5}$ para valores pequeños de PM_{10}.

El modelo RTO ajustado a estos datos es $\boldsymbol{\hat{y}_i = 0.3211 x_i}$ que implica que el estimador de la fracción $PM_{2.5}/PM_{10}$ en la estación Compartir es 0.3211.

Ahora bien, el coeficiente de determinación para este modelo es $\boldsymbol{R^2 = 0.9186}$, que es mucho mayor que el del modelo con intercepto (0.6936). Pero estos dos valores no son comparables porque se obtienen de manera diferente [5], lo que implica la necesidad de acudir a estrategias de comparación

distintas. Siguiendo la recomendación de Rossiter en 2016, se comparan los AIC para ambos modelos, que resultan ser 19,641 para el modelo con intercepto y 19,741 para el modelo sin intercepto, indicando que el modelo con intercepto es mejor [18]. Ahora bien, la principal implicación sobre los estimadores es que en el modelo sin intercepto el estimador de $\boldsymbol{\beta}$ no es insesgado. Pero el sesgo resulta ser menor a medida que el tamaño de muestra es mayor, por lo que con más de 3,000 datos el sesgo es muy cercano a cero. Y dado que el cambio relativo en los AIC es pequeño, se preferirá el modelo sin intercepto. En cuanto a la idea de Casella en 1983, su propuesta se ve muy afectada por el hecho que el dato adicional depende de la cantidad $\boldsymbol{n}^*$, que diverge cuando *n* crece [2].

Nótese que el modelo sin intercepto mejora la estimación para valores pequeños de PM_{10}, en el sentido que no estima valores negativos. Pero para valores altos de PM_{10}, el modelo sin intercepto introduce una subestimación de los valores de $PM_{2.5}$, como se ilustra en la Fig. 2. En defensa del modelo sin intercepto, los valores altos de PM_{10} son menos frecuentes.

3.3. Un modelo aditivo generalizado

Para buscar una solución a esta disyuntiva de los modelos lineales se propone el uso de un modelo aditivo generalizado GAM, que es un modelo de regresión no-paramétrica y en consecuencia no presupone una forma de la función de regresión si no que se estima a partir de las observaciones. En la Fig. 3 se observa que el modelo GAM estima bien la respuesta para valores pequeños y grandes de la variable de predicción, tiene una capacidad explicativa de un 70.3% y tiene un AIC más pequeño que los dos modelos lineales. Sin embargo, el modelo no estima la fracción $PM_{2.5}/PM_{10}$, al menos no de manera directa, aunque sí permite estimar los niveles de $PM_{2.5}$ a partir de los niveles horarios de PM_{10}, lo que permitiría proponer un valor para la fracción, cuyas propiedades deberían estudiarse.

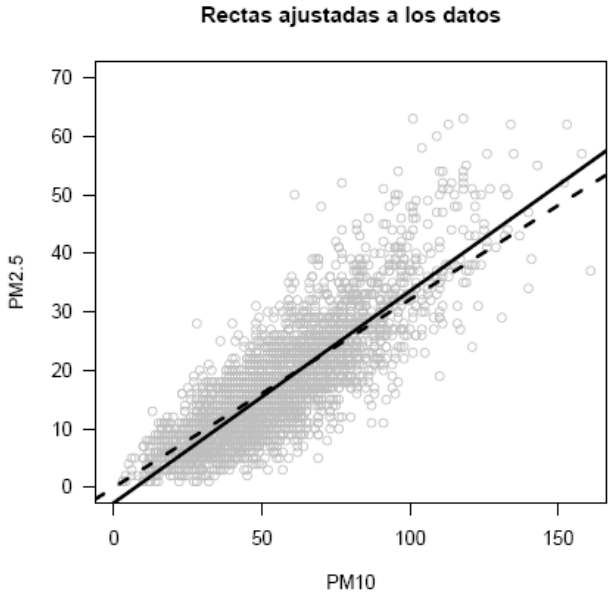

Figura 2. Modelo con intercepto (línea continua) y sin intercepto (línea punteada).
Fuente: Elaboración Propia.

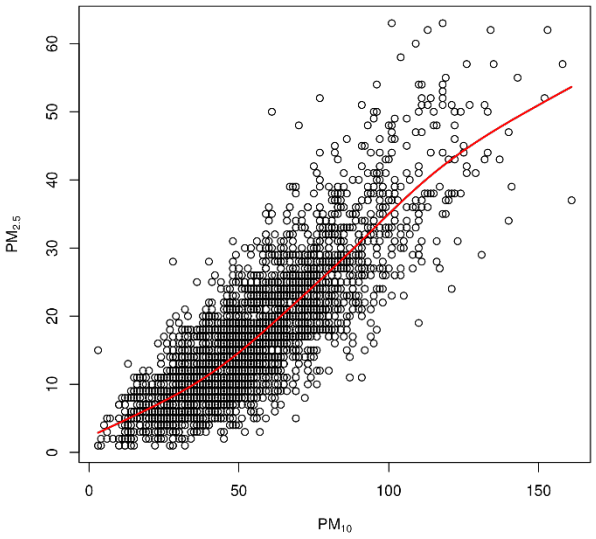

Figura 3. Ajuste GAM.
Fuente: Elaboración Propia.

Una limitación del modelo es que debido a la presencia de una función de suavización en el modelo GAM, no será posible usarlo en estaciones en las cuales no se midan ambos contaminantes, lo que impide la predicción de concentraciones de $PM_{2.5}$ cuando solo se cuenta con mediciones de PM_{10}. Pero definitivamente es una muy buena herramienta cuando se miden ambos contaminantes, por ejemplo para imputar datos faltantes.

3.4. Un modelo de mínimos cuadrados ponderados

Una primera mirada al cumplimiento de los supuestos de los modelos ajustados sugiere un posible crecimiento de la varianza asociado con el crecimiento de los niveles de $PM_{2.5}$. Una solución común en estas situaciones es utilizar un modelo en el cual se introduzca algún tipo de ponderaciones. Como los residuales tienen media cero pero no son homocedásticos, aún si los errores lo son, una idea de uso común es usar como pesos los cuadrados de los residuales. Se ajusta entonces un modelo de este tipo, que conduce a la estimación $\hat{y}_i = -2.623 + 0.3615x_i$. Es decir, los estimadores de los coeficientes de regresión son casi idénticos a los del modelo inicial con intercepto. En este caso el $\boldsymbol{R^2}$ crece hasta 99.95% y el AIC es considerablemente menor que el de los otros modelos. Al igual que el modelo original, este modelo producirá estimaciones negativas para valores pequeños de PM_{10} y tiene la misma dificultad que el modelo GAM, en el sentido que solamente sería útil en los casos en los cuales se miden ambos contaminantes en la misma estación, porque de otra forma no hay manera de calcular los residuales. Una nota de precaución es necesaria en este punto, porque se presentan residuales al cuadrado muy altos que implican el uso de pesos muy pequeños que tienen un gran impacto sobre la matriz de covarianzas, afectando severamente la estructura del error. Y además los supuestos del modelo no se cumplen, en particular el supuesto de normalidad.

3.5. Modelos entre estaciones

Teniendo en cuenta el contexto de contaminación en el que se está trabajando, es posible que las estaciones que se encuentran relativamente cercanas o que presenten condiciones similares con respecto a las actividades antropogénicas, muestren también similaridad en las mediciones de los contaminantes de interés. Por tal motivo, en la Fig. 4 (izquierda) se muestra una comparación entre la relación de PM_{10} y el $PM_{2.5}$ dentro de la estación Compartir, entre tanto, en la Fig. 4 (derecha) se muestra una relación entre estaciones usando el PM_{10} de la estación transitoria como variable de predicción y el $PM_{2.5}$ de la estación compartir como variable de respuesta. Estas estaciones se encuentran a una distancia lineal aproximada de 3.7 km y en zonas con grandes similaridades desde el punto de vista de la generación de contaminantes ambientales. La gráfica apoya la posibilidad de ajustar un modelo lineal basado en esta relación entre estaciones aludiendo semejanzas ambientales y de cercanía espacial.

Con base en lo anterior, se propone el modelo tomando como variable respuesta el $PM_{2.5}$ de la estación Compartir y como variable explicativa el PM_{10} de la estación Transitoria. En el ajuste por mínimos cuadrados sin intercepto se estima la fracción $PM_{2.5}/PM_{10}= 0.3441$, con un $\boldsymbol{R^2} = 0.8145$.

Salta a la vista la similaridad entre los coeficientes obtenidos, siendo 0.321 para el modelo dentro de la estación y 0.344 para el modelo entre estaciones, lo cual reafirma la similitud de las condiciones en ambos sectores. Siguiendo esta misma idea, se procede a ajustar varios modelos entre estaciones similares. Algunos resultados de interés se resumen en la Tabla 1. No se incluyen las estaciones Base Aérea y ERA, porque a pesar de su cercanía geográfica se trata de estaciones con comportamientos ambientales diferentes. Mientras la base aérea está rodeada de una zona industrial, aparte del impacto de la quema de combustible propio de la operación aérea, la Escuela República Argentina está en una zona residencial. El flujo vehicular en ambas estaciones es similar.

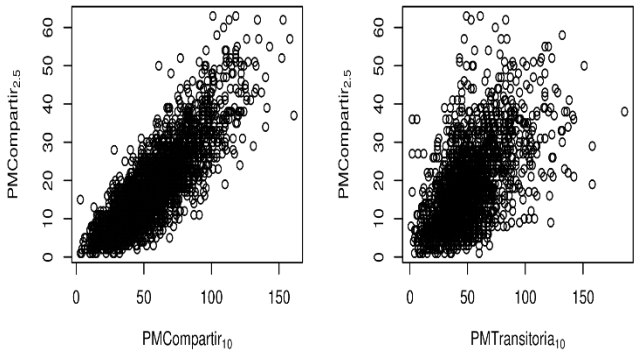

Figura 4. Comparativos
Fuente: Elaboración propia.

Tabla 1.
Resultados modelos cruzados.

Estación PM_{10}	Estación $PM_{2.5}$	Distancia entre Estaciones	Coeficiente Estimado	R^2
Transitoria	Compartir	3.7 km	0.344	0.815
Pance	Univalle	10 km	0.448	0.718
Cañaveralejo	Univalle	4.7 km	0.416	0.807

Fuente: Cálculos propios.

4. Conclusiones, recomendaciones y trabajo futuro

El modelo lineal general (GLM), que ajusta una recta con intercepto, no parece ser la mejor solución para estimar la fracción $PM_{2.5}/PM_{10}$, en primer lugar porque la pendiente de esta recta no podría asimilarse a una fracción; y, en segundo lugar, por su incapacidad para producir valores positivos cerca del origen.

Por su parte, el modelo aditivo generalizado (GAM) permite superar los problemas de estimación para valores pequeños de los contaminantes. Pero formalmente no produce una estimación de la fracción $PM_{2.5}/PM_{10}$. Su uso se vería limitado a una posible estrategia de imputación de datos faltantes en las estaciones en las que se midan ambos contaminantes.

El modelo lineal mediante mínimos cuadrados ponderados (WLS) debe mirarse con cuidado en este problema, en la medida que el supuesto de normalidad de los errores no se satisface, con las implicaciones conocidas sobre la calidad del estimador.

El modelo lineal general a través del origen (RTO), que no ajusta un intercepto, luce como la alternativa más adecuada para estimar la fracción $PM_{2.5}/PM_{10}$, ya que las propiedades de este estimador se derivarían de las propiedades del estimador de regresión, una vez verificado el cumplimiento de los supuestos del modelo.

El uso de la metodología de modelos lineales entre estaciones parece ser coherente en el contexto de contaminantes aéreos siempre y cuando se presente similaridad de condiciones y cercanía entre las estaciones, dado que en estos casos los estimadores de los parámetros son similares. Por tal motivo se propone como una alternativa para proponer una estimación de la concentración horaria de partículas finas en el aire en sitios en los cuales se mide solo un contaminante, ya sea $PM_{2.5}$ o PM_{10}.

Algunas opciones analíticas que podrían aún explorarse en este problema podrían ser la regresión cuantílica, posiblemente alguna variación robusta, y la aplicación de modelos basados en análisis de datos funcionales.

Agradecimientos

Este artículo es uno de los resultados del proyecto de investigación “Modelación estadística de la contaminación del aire por partículas de diámetro aerodinámico menor que 2.5 µm ($PM_{2.5}$)” (CI 2842), que fue financiado por la Vicerrectoría de Investigaciones de la Universidad el Valle.

Referencias

[1] Agresti A., Foundations of linear and generalized linear models, Wiley, 2015.

[2] Casella, G., Leverage and regression through the origin, The American Statistician. [online]. 37(2), pp. 147-152, 1983. Available at: http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1983.10482728

[3] Departamento Administrativo de Gestión del Medio Ambiente, Informe de diseño y operación del sistema de vigilancia de la calidad del aire, DAGMA, Cali, Colombia. [online]. 2012. Available at: www.cali.gov.co/descargar.php?idFile=7672.

[4] Echeverri, C. y Maya, G., Relación entre las partículas finas ($PM_{2.5}$) y respirables (PM_{10}) en la ciudad de Medellín, Revista Ingenierías Universidad de Medellín. [en línea]. 7(12), pp. 23-42, 2008. Disponible en: http://www.redalyc.org/html/750/75011517002/

[5] Eisenhauer, J.G., Regression through the origin, Teaching Statistics, 25, pp. 76-80, 2003.
[6] Environmental Protection Department, Guidelines on the estimation of $PM_{2.5}$ for air quality assessment in Hong Kong, EPD, Atmospheric Environment. [online]. 40, pp. 2735-2749, [date of reference April 30th of 2017]. Available at: http://www.epd.gov.hk/epd/english/environmentinhk/air/guide_ref/guide_aqa_model_g5.html.
[7] Environmental Protection Agency, Particulate matter, EPA, United States, [online]. 2015. Available at: www3.epa.gov/pm/.
[8] Fox, J., Applied regression analysis and generalized linear models, 3era edition, SAGE, 2016.
[9] Gomiscek, B., Hauck, H., Stopper, S. and Preining, O., Spatial and temporal variations of PM_1, $PM_{2.5}$, PM_{10} and particle number concentration during the AUPHEP-project, Atmospheric Environment, 38, pp. 3917-3934, 2004.
[10] Hastie, T.J. and Tibshirani, R.J, Generalized additive models, Chapman & Hall/CRC, USA, 1990.
[11] Munir, S., Analysing temporal trends in the ratios of $PM_{2.5}/PM_{10}$ in the UK, Aerosol and Air Quality Research, 17, pp. 34-38, 2017.
[12] Neter, J., Wasserman, W. and Kutner, M., Applied linear regression models, Richard D. Irwin, INC., 1983.
[13] Organización Mundial de la Salud, Calidad del aire ambiente (exterior) y salud, OMS, [en línea]. 2016. Disponible en: www.who.int/mediacentre/factsheets/fs313/es/.
[14] Othman, S.A., Comparison between models with and without intercept, General Mathematics Notes, 21, pp. 118-127, 2014.
[15] Pace, T.G., Examination of the multiplier used to estimate $PM_{2.5}$ fugitive dust emissions from PM_{10}, 2014.
[16] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, [online]. 2017, Available at: https://www.R-project.org/.
[17] Rojas, N. y Galvis, B., Relación entre $PM_{2.5}$ y PM_{10} en la ciudad de Bogotá, Revista de Ingeniería Universidad de los Andes 22, pp. 54-60, 2005.
[18] Rossiter, D.G., Technical note: Curve fitting with the r environment for statistical computing, 2016.
[19] Smyth, S.C., Jiang, W., Yin, D., Roth, H. and Giroux, E., Evaluation of CMAQ O_3 and $PM_{2.5}$ performance using Pacific 2001 measurement data, Atmospheric Environment, 40, pp 2735-2749, 2006.
[20] Wood, S.N., Generalized additive models: An introduction with R, Chapman & Hall-CRC, 2006.

J. Olaya-Ochoa, es Tecnólogo Químico en 1977, Estadístico 1985 de la Universidad del Valle, MSc Ciencias Matemáticas 1997, PhD Management Science 2000, Clemson University. Profesor titular Universidad del Valle.
ORCID: 0000-0001-7014-2782

D.P. Ovalle-Muñoz, es Estadística de la Universidad del Valle en 2014, actualmente estudiante de Maestría en Estadística de la Universidad del Valle.
ORCID: 0000-0002-5408-5762
C.L. Urbano-León, es Matemático de la Universidad del Cauca en 2012, actualmente estudiante de Maestría en Estadística de la Universidad del Valle.
ORCID: 0000-0003-4622-538X