Vol. 19, n.° 2, pp. 1-10, 2020

Revista UIS Ingenierías

Página de la revista: revistas.uis.edu.co/index.php/revistauisingenierias

Imputación de valores perdidos y detección de valores atípicos en datos funcionales: una aplicación con datos de PM_{10}

Missing value imputation and outlier detection for functional data: an application for PM_{10} data

Rafael Meléndez[1], Stevenson Bolívar [2], Roberto Rojano [3]

[1]Grupo de Universidad Paralela, Facultad de Ingeniería, Universidad de La Guajira, Colombia. Orcid: 0001-7203-1015. Correo electrónico: rmelendez@uniguajira.edu.co

[2] Facultad de Ingeniería, Pontificia Universidad Javeriana, Colombia. Orcid: 0000-0002-5811-2027. Correo electrónico: s_bolivar@javeriana.edu.co

[3] Grupo GISA, Facultad de Ingeniería, Universidad de La Guajira, Colombia. Orcid: 0002-2380-4840. Correo electrónico: rrojano@uniguajira.edu.co

Recibido: 19 septiembre, 2019. Aceptado: 15 febrero, 2020. Versión final: 5 marzo, 2020.

Resumen

Los datos recopilados en el monitoreo de la contaminación del aire, como PM_{10}, se obtienen en estaciones automatizadas que generalmente contenían valores faltantes debido a fallas de la máquina, mantenimiento de rutina o errores humanos. Los conjuntos de datos incompletos pueden causar sesgo de información, por lo tanto, es importante encontrar la mejor manera de estimar estos valores faltantes para garantizar la calidad de los datos analizados. En este trabajo se evaluaron los datos de partículas PM_{10} consideradas en el tiempo como un objeto funcional, para este caso se utilizó la base de datos de la red de monitoreo ambiental de la Corporación Ambiental de La Guajira (Corpoguajira). En este estudio hemos implementado la metodología de Jeng-Min Chiou, (2014) para imputar datos funcionales. La detección de valores atípicos de contaminantes es muy importante para el monitoreo y control de la calidad del aire. Además, hemos implementado el método de imputación de datos faltantes y detección de valores atípicos para datos funcionales. Consideramos las concentraciones de partículas PM_{10} en las estaciones de monitoreo ambiental sobre el área de influencia de la mina de carbón a cielo abierto durante 2012. Para imputar datos faltantes funcionales, se basó en la aplicación de herramientas como el análisis de componentes principales funcional (ACPF) y los procedimientos gráficos para detectar curvas de valores atípicos como el bagplot funcional y el diagrama de caja funcional de la región de mayor densidad (HDR) por sus siglas en ingles. Los resultados indican que la estación de Barranca es una curva atípica y se observó que los intervalos imputados capturan la dinámica que se comparte con las otras trayectorias de las diferentes estaciones.

Palabras clave: datos funcionales; análisis de componentes principales funcionales; valores atípicos funcionales; particulado PM_{10}.

Abstract

The data collected in the air pollution monitoring such as PM_{10} is obtained at automated stations that generally contained missing values due to machine failures, routine maintenance, or human errors. Incomplete data sets may cause information bias. Therefore, it is important to find the best way to estimate these missing values to ensure the

ISSN impreso: 1657 - 4583. ISSN en línea: 2145 - 8456,

R. Meléndez, S. Bolívar, R. Rojano, "Imputación de valores perdidos y detección de valores atípicos en datos funcionales: una aplicación con datos de PM_{10}," *Rev. UIS Ing.*, vol. 19, no. 2, pp. 1-10, 2020. doi: 10.18273/revuin.v19n2-2020001

quality of the analyzed data. In this paper PM_{10} particulate data considered in time as a functional object were evaluated, for this case the database of the environmental monitoring network of the Environmental Corporation of La Guajira (Corpoguajira) was used. In this study we have implemented the methodology by Jeng-Min Chiou (2014) to impute functional data. The detection of outliers of pollutants is very important for monitoring and control of air quality. Additionally, we have implemented the method of imputation of missing data and detection of outliers for functional data. We considered PM_{10} particle concentrations in the environmental monitoring stations over the area of influence of the open pit mining during 2012. To impute functional missing data, it was based on applying tools such as functional principal component analysis (ACPF) and graphic procedures to detect outlier curves such as the bagplot and functional highest density region (HDR) boxplot. The results indicate that Barranca station is an atypical curve and it was observed that the imputed intervals capture the dynamics that are shared with the other trajectories of the different stations.

Keywords: functional data; principal component analysis functional; functional outliers; particulate PM_{10}.

1. Introducción

El manejo y análisis de los datos de calidad de aire con idoneidad facilitan comprobar el cumplimiento de las normas de calidad y evaluar la exposición de la población a los contaminantes del aire [1, 2]. Por ejemplo, la información en el monitorio de particulado generado en las zonas minería es de gran relevancia ambiental y para la salud de los habitantes cercanos a estas zonas por la fuerte evidencia de efectos adversos como la disminución de la función pulmonar, dificultad respiratoria e hipoxemia, cambios adversos en la función autonómica cardíaca [3], [4] y [5].

La mina de carbón del Cerrejón, ubicada en el departamento de La Guajira al norte de Colombia, representa una de las minas más grandes a cielo abierto del mundo, con aporte del 38% en la producción y más del 50% en la exportación, convirtiéndose en la principal compañía de obtención de carbón térmico en el país [6]. Esta capacidad de producción hace que las operaciones mineras a cielo abierto realizadas generen material particulado en cantidades significativas. La masa de partículas menor de 10 μm (PM_{10}) es uno de los contaminantes de referencias más importantes que se utilizan para evaluar la calidad del aire en Colombia.

En Colombia a través de la Resolución 650 de 2010, por la cual se adopta el Protocolo para el Monitoreo y Seguimiento de la Calidad del Aire, establece Sistemas Especiales de Vigilancia de Calidad de Aire (SEVCA) a cualquier población con problemáticas específicas de calidad del aire como minería y alto nivel de industrialización. La Corporación Ambiental de La Guajira (Corpoguajira) cuenta con una red de diez estaciones (10) ubicada en la zona de influencia de la Mina el Cerrejón. Estas estaciones registran mediciones de partícula sedimentable, TSP, PM_{10} y $PM_{2.5}$.

En la practica el análisis de datos generados por particulado en zonas carboníferas de cielo abierto es realizado con métodos estadísticos tradicionales univariado y multivariados, sin tener en cuenta el comportamiento en el tiempo y sin considérala como una trayectoria. La importancia de darle el enfoque funcional o de que las observaciones sean curvas, es que el objeto de estudio (la estación de monitoreo) es considerada como una unidad total en un periodo determinado de tiempo (año). A partir de considerar cada estación como curvas, se realiza el analice y tratamiento que dé a lugar.

El monitoreo de la calidad del aire se realiza con el fin de detectar cualquier concentración significativa de contaminante que puede tener posibles efectos adversos para la salud humana. Sin embargo, dicho análisis se complica si existen gran cantidad de observaciones faltantes, que pueden ser originado por diferentes motivos [7]. Los conjuntos de datos incompletos pueden causar sesgo, por lo tanto, es importante encontrar la mejor manera de estimar valores faltantes que garanticen que los datos analizados presenten información de alta calidad [8].

En la literatura del ADF se ha propuesto diversas técnicas para imputar datos, por ejemplo, [8] compara el método de interpolación lineal y el método de remplazar el dato por la media en un estudio de particulados PM_{10}.

Además, en el caso de datos de precipitación y modelos hidrológicos, generalmente, se aplican métodos de regresión o un método de distancia ponderado que son los más comunes para estimar los datos faltantes en esto temas [9]. Mientras en problemas de particulado de aire se aplican métodos como, [10] y [11], donde se emplearon técnicas simples, como la "*mean top bottom*" para remplazar los valores faltantes en un conjunto de concentraciones de PM_{10} [11]. Sin embargo, en [11] se encontró que este método presenta buenos resultados sólo cuando el número de los valores faltantes es pequeño. Mientras tanto, [12] aplica el método de imputación del vecino más cercano para estos casos.

Los diferentes métodos de imputación presentan sus ventajas y desventajas en cada escenario de trabajo. Es por esto que distintos procedimientos podrían dar lugar a resultados desiguales.

Una de las maneras de hacer control sobre la generación de particulados de aire en el monitoreo ambiental es a través de técnicas de identificación de valores atípicos, que hasta hace poco se basaban en herramientas de detección univariados o bivariados [13], [14] y [15].

La detección de valores atípicos es un campo muy amplio, que se ha estudiado en diferentes contextos. [16], [17] y [18], además de proporcionar una amplia descripción de las diferentes técnicas, se ha examinado en distintos dominios de datos, incluidos los de alta dimensión [19] y en series de tiempo [20], [21].

La detección de datos irregulares es muy popular en estudios de la industria y monitoreo ambiental y, por lo tanto, se han creado algunas herramientas en software para la detección eficiente de estos valores, como el software R (paquetes "outliers'1 y" outlierD " [22]).

En el ADF se ha implementado algunas herramientas para detectar valores extremos, como por ejemplo [23] en estudios sobre calidad del agua y [24] sobre valores atípicos en la calidad del aire.

Adicionalmente, la detección de valores atípicos es otro tema importante en la investigación y análisis de calidad del aire, sobre todo en el contexto funcional. Esto incluyen la detección de atípicos temporales en términos de magnitud e identificación de patrones inusuales (es decir, curvas atípicas de forma), que también proporciona información muy útil a través de sus métodos gráficos [25]. Una visión más global en el caso funcional sobre temas de los datos faltantes se puede encontrar en [26] y [27].

En el ADF existen diverso enfoque para abordar el problema de datos faltantes, entre las técnicas implementadas en la última década están los autores [26], [27]. Además, existe una gran cantidad de literatura que aborda algunos métodos para imputar datos, en el caso multivariado [28] y para datos longitudinales esta [29]. Recientemente, [30] propuso abordar estos temas utilizando el análisis probabilístico de componentes principales (ACPP) y el análisis de componentes principales bayesiano (ACPB) para los algoritmos de imputación.

En cuanto a los métodos de detección de valores atípicos, un punto esencial son las observaciones periféricas. Mientras la detección de datos multivariados atípicos se ha desarrollado a lo largo de varias décadas, pero en observaciones funcionales sólo ha sido discutida en los últimos años [30] y [31]. Es de aclarar que observaciones atípicas pueden conducir de manera adversa a errores de modelación, estimación de parámetros y resultados incorrectos.

Los mecanismos de detección de atípicos en el caso funcional incluyen el uso de análisis de componentes principales robusta [32], parcelas de descomposición de valores singulares [33] y métodos gráficos como el rainbow plot, bagplot y boxplot para datos funcionales [25]. Hace dos décadas el ADF introdujo algunos métodos estadísticos para estudiar este tipo de datos y ha sido ampliamente desarrollados por [25] y [26].

En este trabajo se proponen el análisis de datos funcionales al considerar las concentraciones semanales de PM_{10} como una trayectoria en el tiempo, este es su primer aporte. El enfoque de imputar datos funcionales es muy resiente y de gran relevancia porque es muy común encontrar observaciones faltantes en estudios de monitoreo ambiental. La metodología propuesta para imputar curvas se basa en determinar la esperanza condicional del ACPF, ver (4) y (5), para obtener los escores generados por este mismo y así, ser utilizado como elementos para imputar datos faltantes en el análisis de concentraciones de PM_{10} de las diferentes estaciones de muestreó. Para esto se ha utilizado la metodología expuesta por [34] y obtener la muestra completa.

2. Metodología

Es muy común encontrar datos faltantes en procesos de observación, sobre todos cuando se trata de monitoreo ambiental. Existen muchos motivos entre ellas fallas de equipos, problemas con la electricidad, errores humanos entre otros. Esto nos coloca en un primer escenario de dificultades. Incluso, el ADF tampoco escapa a esta situación, sin olvidar que si faltan algunas observaciones no es un problema grave, solamente cuando se encuentra intervalos muy grandes se dificultan recuperar la estructura completa de la curva.

La imputación de datos funcionales ha sido muy poca tratada en la literatura, es por esto, que hemos querido adoptar la metodología expuesta por [34] al caso de imputación da datos de particulado cuando la observación es una trayectoria.

El enfoque se basa en recuperar la estructura de las curvas utilizado la herramienta del ACPF y la expansión de Karhunen–Loève KL, esta última expresión (KL) no es

más que la media del proceso más un término de error, la cual es estimada a partir de los escores del ACPF [34].

Inicialmente se describe la ubicación de las estaciones de muestreo y los equipos utilizados para este fin, siguiendo con la metodología de la reconstrucción de las observaciones faltantes para su imputación. Finalmente, se obtiene las muestras completas y a partir de ahí se procede con la búsqueda de curvas atípicas.

2.1. Área de estudio

El Cerrejón es una mina de explotación de carbón a cielo abierto localizada al norte de Colombia en el departamento de La Guajira, frontera con Venezuela. Es una de las minas a cielo abierto más grandes de Latinoamérica y decimo segunda en el mundo. Está ubicada en las coordenadas 11°5'2'' Norte y 72°40'31'' Oeste (Figura 1).

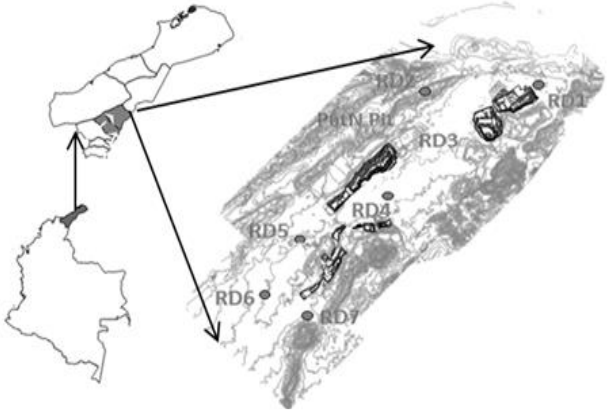

Figura 1. Localización del área de estudio (zona minera a cielo abierto en La Guajira).

Los depósitos de carbón se encuentran en una extensión aproximada de 70.000 hectáreas. La topografía del terreno es representada por una llanura aluvial entre los sistemas montañosos de la Serranía del Perijá por el oriente y la Sierra Nevada de Santa Marta por el occidente.

2.2. Datos de estudios PM_{10}

Se tomaron para el estudio los datos de material particulado PM_{10}, determinados en 7 estaciones que conforman la Red de Calidad de Aire de el Cerrejón. Las concentraciones fueron recolectadas por la empresa entre el 4 de enero y el 29 de diciembre de 2012. Las partículas PM_{10} se determinaron por el método gravimétrico, utilizando muestreadores de alto de volumen (Hi-Vol) con cabezal de separación inercial para las PM_{10}, según lo establecido en el método de referencia para PM_{10} de la Agencia Ambiental de los Estados Unidos EPA. Los equipos Hi-Vol operaron cada seis días con tiempo de muestreo de 24 horas, desde las 12:00 h hasta las 12:00 h del día siguiente. Se tomaron 61 muestras en cada estación. La Tabla 1 presenta la ubicación de las estaciones, los equipos utilizados en el monitoreo y los parámetros medidos en cada estación.

Tabla 1. Ubicación de las estaciones, equipos y parámetros medidos en cada estación de monitoreo de PM_{10}

Estación	Código	Ubicación	Equipo
Estación 1. Sol y Sombra	RD1	11°09' N - 72° 31 W	Hi-Vol (PM_{10})
Estación 2. Vivienda	RD2	11°09' N - 72°36' W	Hi-Vol (PM_{10})
Estación 3. Roche	RD3	11°04' N - 72°38' W	Hi-Vol (PM_{10})
Estación 4. Patilla	RD4	11°32' N - 72°55' W	Hi-Vol (PM_{10})
Estación 5. Provincial	RD5	11°52' N - 72°55' W	Hi-Vol (PM_{10})
Estación 6. Barrancas	RD6	11°31' N - 72°56' W	Hi-Vol (PM_{10})
Estación 7. Casitas	RD7	11°32' N - 72°54' W	Hi-Vol (PM_{10})

Fuente: elaboración propia.

2.3. Modelos de valores perdidos

En general, los datos faltantes pueden ser de naturaleza aleatoria y algunas veces son causados por fallas del detector o las herramientas de medición. Hay tres clases de patrones típicos de datos faltantes [35], [36] para los datos multivariados y para datos longitudinales [36]. Primero cuando los datos faltantes son completamente al azar, segundo, cuando los faltantes presentan un patrón determinado y por últimos una combinación de los dos, pero en el caso funcional este enfoque es diferente.

En el ADF se adoptan los métodos anteriormente expuesto, y se describe las siguientes situaciones que podrían ocurrir [34].

1. Los puntos faltantes PF están aislados o dispersos aleatoriamente en la curva observada.
2. Se presentan intervalos faltantes IF, o pequeños grupos de puntos perdidos aleatoriamente a través de la curva.
3. Se presentan tanto puntos faltantes PF como intervalos faltantes IF en la curva observada.

En nuestro caso se presentaron intervalos grandes de observaciones faltantes que dificultan la obtención de la curva completa.

A continuación, se describe la técnica para reconstruir las curvas a través del enfoque de ACPF.

2.3.1. Análisis de Componente principales funcional ACPF

El análisis de componentes principales (ACP) es un enfoque estándar para la exploración de la variabilidad en datos multivariados. El ACP utiliza una descomposición de valor propio de la matriz de covarianza de los datos para encontrar direcciones en el espacio de observaciones a lo largo de las cuales los datos tienen la mayor variabilidad. Para cada componente principal, el análisis produce un vector de peso que proporciona la dirección de la variabilidad correspondiente a ese componente [37].

En el contexto funcional, cada componente principal está especificado por una función de peso o función propia $\phi(t)$ definida en el mismo rango de $\tau = [0, T]$ que los datos funcionales. De forma similar el ACPF resuelve el problema obteniendo las funciones propias $\phi_i(t), i = 1, \dots, m$, valores propios λ_k y la función de covarianza $Cov_X(s,t) = cov(X(s), X(t)) = G(s,t), s, t \in \tau,$ que en la práctica es estimada a partir de la muestra de datos funcionales. De igual manera esta técnica permite encontrar los escores de cada curva, que se proyectan como un vector de datos en el espacio reducido [38].

La mayoría de los enfoques de datos funcionales son no paramétrico debido a las características de los datos, dado que se requiere unos supuestos mínimos en comparación con el modelo paramétrico. En el caso funcional para imputar se hace uso del ACPF.

Primero , se adopta la idea de que cada trayectoria de particulado PM_{10} es una realización de una función aleatoria $X(t)$. Donde se asume que el proceso estocástico $X(\cdot)$ es la función aleatoria de particulado que tiene una curva media suavizada desconocida $EX(t) = \mu(t)$ con función de covarianza definida, así $cov(X(s), X(t)) = G(s,t), s, t \in \tau$, donde $\tau = [0, T]$ se encuentra en el espacio cuadrado integrable L^2. Aquí G tiene una expansión ortogonal en L^2. La solución del problema resulta de resolver la siguiente ecuación [39].

$$\int Cov_X(s,t)\phi(t)dt = \lambda\phi(s)\}$$

$G(s,t) = \sum_{i=1}^{\infty} \lambda_k \phi_k(s)\, \phi_k(t)$, así que los $\{\lambda_k\}$ son el conjunto de valores propios y $\{\phi_k\}$ el conjunto de funciones propias que forman un conjunto base ortonormal con norma uno en L^2. Entonces cada trayectoria aleatoria de particulado PM_{10} tiene la siguiente expansión Karhunen - Loéve.

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t). \quad (1)$$

En la práctica, la función aleatoria X_i está contaminada por errores de medida y la i-ésima curva puede expresarse como:

$$Y_i(t_{ij}) = X_i(t_{ij}) + \varepsilon_{ij}, \quad (2)$$

$$= \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t). \quad (3)$$

Donde los errores de medición aleatorios ε_{ij} se asumen no correlacionados entre sí e independientes de ξ_{ik} con $E(\xi_{ik}) = 0$ y $var(\xi_{ik}) = \sigma^2$. Para obtener la función estimada (1), se utiliza la función del modelo de componentes μ y ϕ_k aplicando localmente el método suavizado mínimos cuadrados ponderado con la combinación de todas las trayectorias para estimar la función media. Para obtener la estimación de G, se suaviza las covarianzas empíricas. Seguidamente se obtiene los λ_k valores propios estimados y los ϕ_k aplicando el procedimiento de la descomposición espectral para la estimación de la función de covarianza en el ACPF, similarmente como se procede en el ACP tradicional.

La estimación de coeficiente aleatorio de ξ_{ik} no se puede obtener fácilmente a través de esta aproximación

$$\hat{\xi}_{ik} = \int \big(X_i(t) - \hat{\mu}(t)\big)\hat{\phi}_k(t)dt$$

Primero, cuando faltan muchas observaciones esta integral de aproximación se dificultad. Segundo, $X_i(t)$ no se puede observar directamente, sino solo a través Y_i que están afectadas por los errores de medición, y estimar ξ_{ik} sustituyendo Y_i por X_i puede conducir a escores sesgados en el ACPF. Para superar estas dificultades, se adopta el enfoque de [38]. Sea Σ_{Y_i} la matriz de covarianza de Y_i , donde los escores ξ_{ik} y los errores ε_{ik} cumplen conjuntamente el supuesto Gaussiano.

$$E(\xi_{ik} | Y_i) = \lambda_k \phi_{ik}^T \Sigma_{ik}^T (Y_i - \mu_i) \quad (4)$$

La estimación condicional de los escores de la ecuación (33) son obtenidos de la siguiente manera:

$$\hat{\xi}_{ik} = \hat{\xi}_{ik}\hat{\phi}_{ik}^T \hat{\Sigma}_{ik}^T (Y_i - \hat{\mu}_i) \quad (5)$$

Donde, la matriz de covarianza de Y_i es estimada por $\hat{\Sigma}_{Y_i} = \{\hat{G}(t_{ij}, t_{il}) + \hat{\sigma}^2\delta_{ij}\}_{1 \le j,l \le m_i}$ y δ_{ij} es el delta de

kronocker. Para la estimación de valores perdidos se utilizaría la metodología anteriormente detallada.

A continuación, se aplica el criterio de la fracción de la varianza explicada con la regla de los primeros L componentes que explican al menos $\tau_\lambda \times 100\%$ de la varianza total.

$$L = min\left\{\frac{\sum_{k=1}^{L}\hat{\lambda}_k}{\sum_{k=1}^{M}\hat{\lambda}_k} > \tau_\lambda\right\}, \quad (6)$$

donde M es el número mayor de componentes con $\hat{\lambda}_k > 0$ y τ_λ es un valor predeterminado.

Basado en la estimación de un modelo de componente $\hat{\mu}$,$\hat{\phi}$ y $\{\hat{\xi}_{ik}\}_{k=1,\ldots,L}$ para todo *i*, la predicción de la función de la *i-ésima* curva, está dada por

$$\hat{X}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{L}\hat{\xi}_{ik}\hat{\phi}_k(t) \quad (7)$$

Los datos ajustados del modelo (4) se pueden utilizar para imputar una observación faltante por el valor $\hat{X}_i(t_{ij})$. Observe que las trayectorias pronosticadas incluyen las componentes de la función media suavizada y una combinación lineal de las funciones propias, que recuperar las trayectorias individuales de las mediciones del ruido, mientras que el error depende del número de componentes seleccionados en el modelo ACPF.

A partir de la muestra de observaciones completas, se continua con la búsqueda de valores atípicos que podría entregar algún indicio que las concentraciones de PM_{10} sobrepasan los estándares establecidos por la ley ambiental o presentar una alarma al respecto.

A continuación, se describe como se implementó estas herramientas gráficas para la búsqueda de atípicos funcionales en nuestro estudio.

2.3.2. Herramienta gráficas para detectar atípicos funcionales

Las herramientas gráficas para obtener valores atípicos en el caso funcional son de mucha utilidad. Entre los más utilizados están las gráficas bagplot funcional y el HDR boxplot funcional, propuesta por [25]. Ambos métodos se basan en tomar los dos primeros escores de los componentes principales del ACPF. Para este propósito, [36] aplica un algoritmo robusto que permite intuir que las proyecciones desde la descomposición de componentes principales deben ser sensibles a valores atípicos. Sin embargo, este método no considera valores perdidos y se requiere un conjunto de datos completos para proyectar los datos en un espacio de dimensión menor de tal manera que una medida robusta de la varianza de los datos proyectados sea máxima.

Como primer paso se imputaron las trayectorias de PM_{10} que presentaban intervalos de faltantes muy grandes, implementando la metodología desarrollada por [34]. Con la muestra completa de datos se aplicó los métodos gráficos HDR boxplot y bagplot funcional con la idea de encontrar algunos valores atípicos.

La herramienta grafica consiste en aplicar ACPF a las curvas reconstruidas, que en este caso corresponde a las trayectorias de las diferentes estaciones de monitoreo de PM_{10}, luego con los dos primeros escores del ACPF se construyen regiones bivariadas. Los métodos pretenden construir una bolsa que contenga el 95% de los escores y los puntos que este por fuera se ponderan como valores atípicos.

2.3.3. Detección de valores atípicos funcionales

Para, llevar a cabo el procedimiento en la detección de valores atípicos en los gráficos funcionales, se necesita pre especificar una probabilidad de cobertura para determinar la región periférica. La probabilidad de cobertura generalmente suele ser de 99%, 95% y 90%, correspondiente a un nivel de significancia de $\alpha = 0.01, 0.05\ y\ 0.10$.

3. Resultados

Los resultados de la Figura 2 no evidencia diferencia significativa entre los promedios de las siete (7) estaciones. Se observa mayores concentraciones en la estación RD7 (Casitas) y la menor concentración en la estación RD1 (Sol y Sombra), la cual está ubicada viento arriba de la zona de explotación. El rango de las concentraciones de PM_{10} para el periodo de estudio varió entre 2,00 μg/m^3 y 87 μg/m^3. La estación RD7 reportó el mayor promedio en las concentraciones de PM_{10} con un nivel de 41 μg/m^3 y RD1 presentó el menor promedio con un nivel de 17 μg/m^3.

Los resultados muestran que el material particulado PM_{10} son inferiores a los estándares nacionales y de la EPA, a pesar de estar en una zona de influencia minera. Un reciente proyecto de la EPA sugiere nuevo estándar de 24 horas para PM_{10} en el rango de 65 - 85 μg/m3. Aun para los niveles proyectados se estarían cumpliendo estos preceptos.

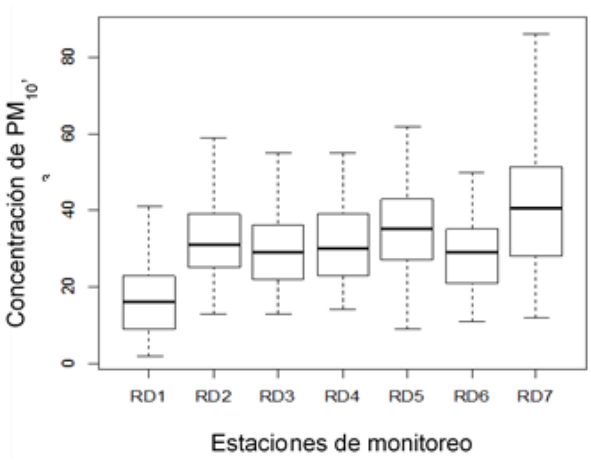

Figura 2. Variabilidad en las concentraciones de PM_{10} en las estaciones de muestreo.

En la Figura 3 y Figura 4 ilustra el HDR boxplot y bagplot funcional, donde la estrella negra corresponde a la curva mediana, mientras la región gris oscura pertenece al 50% de las curvas y la región gris claro corresponde al 95% de ellas. Mientras que, el punto 10 (Figura 3) representa a la estación de monitoreo (Provicial) que se muestra como una trayectoria atípica a un nivel de significancia de $\alpha = 0.05$. Esta curva presenta los niveles más bajos de concentración a lo largo del tiempo de estudio, lo que en este caso es muy positivo. Mientras que la Figura 4 muestra el bagplot funcional, donde la estrella representa la curva mediana y la región gris clara corresponde al 95% de las curvas. Este grafico muestra como atípica la estación de monitoreo Barranca, debido a que presenta los niveles de concentración más alta a lo largo del tiempo de estudio.

La estación Barrancas se presenta como atípica dado que las variaciones de concentraciones son las más altas, con respecto a la curva mediana. A pesar de estar en el rango permitidos por la norma ambiental. Aunque esta curva es atípica es de forma muestra sus picos más altos cerca de los meses de septiembre y octubre.

Los resultados muestran que la estación (Barrancas) es una curva atípica, lo que indica que se debe proponer estudios más profundos y monitoreo más seguido en el tiempo sobre esta zona, que permita una evaluación constante.

En diferentes sitios de muestreo los resultados de las concentraciones de PM_{10} no excedieron los estándares diarios de la Norma Colombiana y la National Ambient Air Quality Standards (NAAQS) de los Estados Unidos.

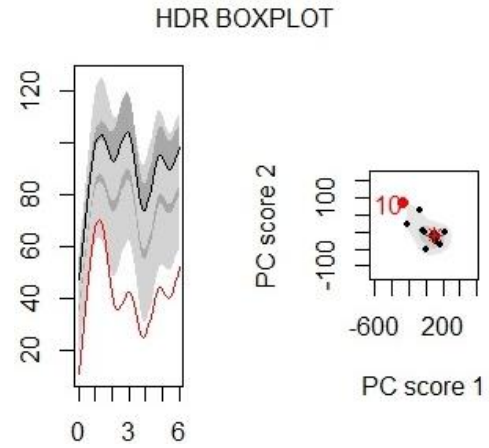

Figura 3. Los gráficos HDR bivariado y bagplot HDR funcional para las concentraciones de PM_{10} de las diferentes estaciones.

La región gris oscuro y gris claro de la Figura 3 muestran el 50% de HDR y el HDR externo, respectivamente. La línea negra es la curva modal. Las curvas fuera de la región exterior son valores atípicos (Provincial).

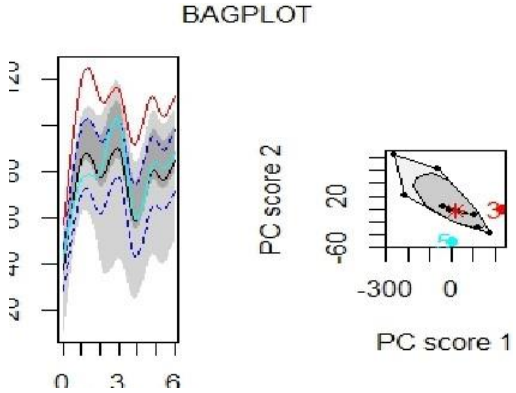

Figura 4. Gráfica de bolsas bivariada y gráfica de bolsas funcional para las concentraciones de PM_{10} de las diferentes estaciones.

Las regiones oscuras y gris claro en la figura 4 muestran la bolsa de regiones. El asterisco es la mediana de profundidad de Tukey. A la izquierda, la línea negra es la curva mediana, rodeada por intervalos de confianza puntuales del 95%. Las curvas fuera de la región exterior se muestran como valores atípicos. (estación Barrancas).

4. Conclusiones

Este artículo presenta un estudio de caso donde se implementa la imputación de curvas en el caso funcional y de detección de atípico en el mismo contexto.

Observaciones faltantes en estudios de monitoreo ambiental son más comunes de lo que se cree, en el

contexto funcional no escapa a este problema, por esto se presentó una metodología que permite reconstruir trayectorias y para ilustrar esta situación, consideramos el caso de trayectorias de concentración PM_{10} a la periferia de una zona de mina de carbón a cielo abierto.

Una de las maneras para evaluar y controlar la calidad del aire es a través de detección de valores atípicos. En el contexto funcional se ha implementado últimamente técnicas graficas que permite obtener curvas atípicas tanto de escala como de forma. Es así que estas curvas pueden suministrar valiosa información para su intervención preventiva en estudios de monitoreo ambiental.

El resultado evidencia que la estación Barrancas como una curva atípica, mostrando niveles de concentración PM_{10} más altas que el promedio, esto es un indicativo para tomar acciones sobre un monitoreo más seguidos y más exhaustivos sobre la calidad del aire. Aunque la curva 10 (Provincial) se presenta como una curva atípica de forma, no debe ser tenida en cuenta para su intervención en un monitoreo más detallado porque presenta bajas concentración de particulado, como es lo deseado.

Nuestros resultados numéricos indican que el enfoque de datos funcionales para la imputación de valores perdidos y la detección de valores atípicos puede aplicarse y ajustarse muy bien a este tipo de estudios.

Agradecimientos

Los autores expresan sus agradecimientos a la Corporación Autónoma Regional de La Guajira por el suministro de datos de la zona de estudio. A Colciencias por el financiamiento y apoyo del proyecto "Aplicación de los modelos de receptor en el aporte de fuentes a la contaminación del aire por material particulado" Código: 1115524304651.

Referencias

[1] J. Duyzer, D. Van Den Hout, P. Zandved & S. Van Ratingen, "Representativeness of air quality monitoring networks," *Atmospheric Environment*, no. 104, pp. 88-101, 2015. doi: 10.1016/j.atmosenv.2014.12.067

[2] L. Zhao, Y. Xie, J. Wang, & J. Xu, "A performance assessment and adjustment program for air quality monitoring networks in Shanghai," *Atmospheric Environment*, no. 122, pp. 382-392, 2015. doi: 10.1016/j.atmosenv.2015.09.069

[3] M. Dostal, A. Pastorkova, S. Rychlik, E. Rychlikova, V. Svecova,, Schallerova, E. & R. Sram, "Comparison of child morbidity in regions of Ostrava, Czech Republic, with different degrees of pollution: a retrospective cohort study," *Environmental Health*, no. 12, pp. 1-11, 2013. doi: 10.1186/1476-069X-12-74

[4] G. Muránszky, M. Ovari, I. Virag, P. Csiba, R. Dobai & G. Zaray, "Chemical characterization of PM10 fractions of urban aerosol," *Microchemical Journal*, no. 98, pp. 1-10, 2011. doi: 10.1016/j.microc.2010.10.002

[5] F. Lu, D. Xu, Y. Cheng, "Systematic review and meta-analysis of the adverse health effects of ambient $PM_{2.5}$ and PM_{10} pollution in the Chinese population," *Environmental Research*, 136, pp. 196–204, 2015. doi: 10.1016/j.envres.2014.06.029

[6] Plan Nacional para el desarrollo minero, Visión año 2019. Ministerio de Minas y Energías, Bogotá, 2012.

[7] N. Noor & M. Zainudin, "A review: Missing values in environmental data sets," *In Proceeding of International Conference on Environment*, 2008 [En línea]. Disponible en: https://scholar.google.com.pk.

[8] N. Noor, M. Abdullah, A. Yahaya, & N. Ramli, "Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set," *Materials Science Forum*, vol. 803, pp. 278-281, 2015. doi: 10.4028/www.scientific.net/MSF.803.278

[9] H. Lee and K. Kang, "Interpolation of Missing Precipitation Data Using Kernel Estimations for Hydrologic Modeling," *Advances in Meteorology*, 2015. doi: 10.4028/www.scientific.net/MSF.803.278

[10] M. Fitri, N. Ramli, A. Yahaya, N. Sansuddin, N. Ghazali, & W. Al Madhoun, "Monsoonal differences and probability distribution of PM_{10} concentration," *Environmental Monitoring Assessment*, vol. 163, pp. 655-667, 2010. doi: 10.1007/s10661-009-0866-0

[11] N. Noor, A. Yahaya, N. Ramli, & M. Abdullah, "The replacement of missing values of continuous air pollution monitoring data using mean top bottom imputation technique," *Journal of Engineering Research & Education*, vol. 3, pp. 96-105, 2006.

[12] N. Shaadan, S. Deni, & A. Jemain, "Assessing and comparing PM10 pollutant behaviour using functional data approach," *Sains Malaysiana*, vol. 41, no. (11), pp. 1335-1344, 2012.

[13] H. Ahn, "Outlier detection in total phosphorus concentration data from South Florida rainfall," *J. Am. Water Resour. Assoc.*, vol. 35, no. 2, pp. 301–310, 1999. doi: 10.1111/j.1752-1688.1999.tb03591.x

[14] R. Gilbert, R. *Statistical methods for environmental pollution monitoring*. Van Nostrand Reinhold, New York, 1987.

[15] K. Reckhow, & S. Chapra, *Engineering approaches for lake management*. Volume 1: Data analysis and empirical modeling, Butterworth, Boston. 1983.

[16] C. C. Aggarwal, *Outlier Analysis*. New York, NY, USA: Springer, 2013.

[17] V. Chandola, A. Banerjee, & V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, 2009.

[18] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev*., vol. 22, no. 2, pp. 85–126, 2004. doi: 10.1007/s10462-004-4304-y

[19] C. C. Aggarwal & P. S. Yu, "Outlier detection for high dimensional data," *SIGMOD Rec*., vol. 30, pp. 37–46, May 2001.

[20] J. P. Burman & M. C. Otto, "Census bureau research project: Outliers in time series," *Bureau of the Census, SRD Res. Rep.*, CENSUS/SRD/RR-88114, May 1988.

[21] A. J. Fox, "Outliers in time series," *J. Roy. Statist. Soc. B Methodol.*, vol. 34, no. 3, pp. 350–363, 1972.

[22] H. Cho, Y. jin Kim, H. J. Jung, S. W. Lee, and J. W. Lee, "OutlierD: An R package for outlier detection using quantile regression on mass spectrometry data," *Bioinformatics*, vol. 24, no. 6, pp. 882–884, 2008.

[23] C. D. Muniz, P. G. Nieto., J. A. Fernandez, J. M. Torres, L. Taboada, "Detection of outliers in water quality monitoring samples using functional data analysis in San Esteban estuary (Northern Spain)," *Science of the Total Environment*, 2012. doi: 10.1016/j.scitotenv.2012.08.083

[24] J. Martinez, A. Saavedra, P. J. Garcia-Nieto, J. L. Piñeiro., C. Iglesias, J. Taboada, J. Sancho, J. Pastor, "Air quality parameters outlier's detection using functional data analysis in the Langreo urban area (Northern Spain)," *Applied Mathematics and Computation*, 2014. doi: 10.1016/j.amc.2014.05.004

[25] R. Hyndman and H. Shang, "Rainbow Plots, Bagplots, and Boxplots for Functional Data," *Journal of Computational and Graphical Statistics, vol.* 19 no. 1, pp. 29–45, 2010.

[26] P. Allison, *Missing Data*. California: Thousand Oaks, Sage, 2001.

[27] J. Schafer & J. Graham, "Missing Data: Our View of the State of the Art," *Psychological Methods*, vol. 7, no. 2, 147–177, 2002.

[28] J. Schafer, *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall, 1997.

[29] E. Beale & R. Little, "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society*, Series B, vol.3, no.1, pp. 129–145, 1975.

[30] L. Qu, L. Li, Y. Zhang, & J. Hu, "PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach," *IEEE Transactions on Intelligent Transportation System*, vol. 10, no. 3, pp. 512–522, 2009. doi: 10.1109/TITS.2009.2026312

[31] C. Chen, J. Kwon, J. Rice, A. Skabardonis & P. Varaiya, "Detecting Errors and Imputing Missing Data for Single-Loop Surveillance System," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 18, no. 55, pp. 160–167, 2003. doi: 10.3141/1855-20

[32] R. Hyndman & M. Ullah, "Robust Forecasting of Mortality and Fertility Rates: A Functional Data Approach," *Computational Statistics and Data Analysis*, vol. 51, no. 10, pp. 4942–4956, 2007. doi: 10.1016/j.csda.2006.07.028

[33] R. Hyndman & H. Shang, "Rainbow Plots, Bagplots, and Boxplots for Functional Data," *Journal of Computational and Graphical Statistics*, vol. 19, no. 1, pp. 29–45, 2010. doi: 10.1198/jcgs.2009.08158

[34] J. Chiou, Y. Zhang, W. Chen & Ch. Chang, "A functional data approach to missing value imputation and outlier detection for traffic flow data," *Transportmetrica B: Transport Dynamics*, vol. 2, no. 2, 106-129, 2014. doi: 10.1080/21680566.2014.892847

[35] D. Rubin, "Inference and Missing Data," *Biometrika*, vol. 63, no. 3, pp. 581–592,1976. doi: 10.2307/2335739

[36] R. Little & D. Rubin, S*tatistical Analysis with Missing Data*, 2nd ed, New York: Wiley, pp. 255-260. 2002.

[37] P. Hall & M. Hosseini-Nasab, "On properties of functional principal components analysis," *J. R. Statist. Soc. B.*, vol. 68, Part 1, pp. 109-126. 2006.

[38] F. Yao, H. G. Müller, & J. L. Wang, "Functional Data Analysis for Sparse Longitudinal Data," *Journal of American Statistical Association*, vol. 100, no. 470, pp. 577–590. 2005. doi: 10.1198/016214504000001745

[39] K. Deregowski & M. Krzysko, "Principal components analysis for functional data," *Colloquium Biometricum*, vol. 41, pp. 5-7. 2011.