

# Imputación, basada en la distribución Normal multivariada, de datos faltantes de mediciones de partículas finas suspendidas en el aire

Imputation, based on the multivariate Normal distribution, of missing records of fine particulate matter in air

E. Arroyave-López  ; A Villarreal-Monsalve  ; J. Olaya-Ochoa 

DOI: <https://doi.org/10.22517/23447214.24734>

Artículo de investigación científica y tecnológica

**Abstract**— We propose and evaluate two imputation methods for missing data of fine particulate matter on air. We assume a 24-variate normal distribution, one per weekday, due to the fact that the imputed data are hourly data on a 24-hour daily cycle. From this distribution properties, the imputation methods are based on the conditional distributions for missing hours, starting from hours with available records. We estimate the weekday variance-covariance matrix using two methods: maximum likelihood (denoted by  $\Sigma$ ), and shrinkage (denoted  $\Sigma^*$ ). Afterwards, we verify the missing completely at random (MCAR) assumption using the Little's test, and also de multivariate normality using the Mardia's test. Finally, we evaluate the proposed methods through a simulation trial, generating suitable scenarios for this kind of problems. We use two evaluation criteria: the coefficient of determination ( $R^2$ ) and the square root of the mean square error ( $RMSE$ ). We use a 2018 data set from Cali, Colombia, to illustrate how to use the proposed methods. We reach  $R^2$  values of around 0,70 and 0,49, and  $RMSE$  values of around 5,7 and 8,5, for the methods based on  $\Sigma$  and  $\Sigma^*$ , respectively.

**Index Terms**—Air pollution; Little's test; Mardia test; Missing points;  $PM_{2,5}$ ;  $R^2$ ;  $RMSE$ ; Shrinkage, Simulation.

**Resumen**— Se proponen y evalúan dos métodos de imputación para datos faltantes de partículas finas suspendidas en el aire, asumiendo que cada día de la semana se puede modelar mediante una distribución normal 24-variada, debido a que los datos que se imputan son datos horarios sobre un ciclo diario de 24 horas. A partir de las propiedades de esta distribución, se conduce la imputación estimando las distribuciones condicionales para las horas faltantes a partir de las horas con información disponible. Para cada día se estima la matriz de varianzas y covarianzas por dos métodos: por máxima verosimilitud (denotada  $\Sigma$ ) y por shrinkage (denotada  $\Sigma^*$ ). Luego, se prueba el supuesto de pérdida

completamente al azar (MCAR) mediante el test de Little y se prueba el supuesto de normalidad multivariada con el test de Mardia. Finalmente, se evalúan los métodos propuestos vía simulación, generando escenarios posibles para este tipo de problemas, junto con dos criterios: coeficiente de determinación ( $R^2$ ) y raíz cuadrada del error cuadrático medio ( $RMSE$ ). Los métodos propuestos se ilustran con datos de mediciones de Cali, Colombia, de 2018. Se alcanzan valores alrededor de 0,70 y 0,49 para el  $R^2$  y de 5,7 y 8,5 para el  $RMSE$ , para los métodos basados en  $\Sigma$  y  $\Sigma^*$ , respectivamente.

**Palabras clave** — Contaminación del aire; Datos faltantes,  $PM_{2,5}$ ;  $R^2$ ,  $RMSE$ ; Shrinkage; Simulación, Test de Little, Test de Mardia

## I. INTRODUCCIÓN

DE acuerdo con la Organización Mundial de la Salud (OMS) [1], la contaminación del aire es el mayor riesgo ambiental para la salud en la actualidad, estimando que contribuye a unos 7 millones de muertes prematuras cada año. Entre los contaminantes atmosféricos se encuentra el  $PM_{2,5}$  el cual consiste en un conjunto de partículas suspendidas en el aire cuyo diámetro es inferior a  $2,5\mu m$ , conformadas, por ejemplo, por polvo, hollín, sal arrastrada por el aire, esporas y humo de incendios forestales [2]. Las principales fuentes de emisión de estas partículas son los incendios forestales, la producción industrial, las emisiones de fuentes móviles y la calefacción, entre otros [3].

Por otra parte, el Observatorio Nacional de Salud de Colombia [4], reporta que en este país la contaminación atmosférica es un factor potencial causante de aproximadamente 15.681 muertes

Este manuscrito fue sometido el 02 de junio de 2021, aceptado el 09 de marzo de 2023 y publicado el 31 de marzo de 2023. Este trabajo contó con el apoyo financiero de la Vicerrectoría de Investigaciones de la Universidad del Valle, proyecto CI-21081

E. Arroyave es Estadístico de la Universidad del Valle y ejerce como analista de datos del Departamento Administrativo de las Tecnologías de y las Comunicaciones (DATIC), en la Alcaldía de Santiago de Cali, Colombia. (e-mail: [esteban.arroyave@correounivalle.edu.co](mailto:esteban.arroyave@correounivalle.edu.co)).

A. Villarreal-Monsalve es Estadístico en la Universidad del Valle, Analista de analítica de datos en Seguridad Atlas LTDA, Cali, Colombia, 760035 (e-mail: [villarreal.alejandro@correounivalle.edu.co](mailto:villarreal.alejandro@correounivalle.edu.co)).

J. Olaya-Ochoa es profesor titular de Estadística en la Escuela de Estadística de la Universidad del Valle, Cali, Colombia, 760031 (e-mail: [javier.olaya@correounivalle.edu.co](mailto:javier.olaya@correounivalle.edu.co)).



anuales asociadas con el sistema respiratorio, el cerebro y enfermedades cardiacas. Además, se ha relacionado al tamaño de las partículas suspendidas en el aire (PM) con el potencial para contraer este tipo de enfermedades. Esto se debe a que las partículas menores a  $2,5 \mu\text{m}$  que ingresan por las vías respiratorias pueden alojarse en los pulmones, o pasar directamente a los vasos sanguíneos [2].

Los datos faltantes en las bases de datos sobre calidad de aire son un problema muy común [5]. Así que en la literatura hay disponible una diversidad de ideas sobre cómo resolver este problema ([6], [7], [8], [9], [10], [11], [12]). En estos trabajos, los datos disponibles son de frecuencia al menos diaria. O sea que si se necesitaran datos con mayor frecuencia, por ejemplo datos horarios, o datos continuos (por ejemplo datos funcionales), habría necesidad de nuevas opciones analíticas. En [13] se evalúan varios métodos de imputación de contaminantes atmosféricos y sugieren adoptar el método de imputar el promedio de los datos faltantes anterior y posterior. Este método no sería muy útil en situaciones en las que se presenten datos faltantes de muchas horas consecutivas.

El estudio de las partículas finas suspendidas en el aire (PM<sub>2,5</sub>), tropieza con el hecho que las mediciones en los Sistemas de Vigilancia de la Calidad del Aire (SVCA) tienen como denominador común la ausencia de muchas mediciones. Esta falta de información, usualmente referida como *datos faltantes*, tiene múltiples orígenes. Por ejemplo, para citar algunas causas comunes, fallas en el suministro de energía eléctrica, fallas en los sistemas de comunicación o situaciones relacionadas con el clima.

En la referencia [14], los autores encuentran que las mediciones de PM<sub>2,5</sub> correspondientes a una hora  $i$  del día  $j$  de la semana, siguen una distribución Normal con media y varianza igual a la media y a la varianza funcionales del día  $j$ , evaluadas en la hora  $i$  de ese día. Estos resultados, con datos de 2015, son validados por [15] con datos de 2017. Así que la distribución de las mediciones de cada hora  $i$  del día  $j$  se puede asumir Normal. En este trabajo se avanza hasta la verificación que la distribución conjunta para las 24 horas del día  $j$  es Normal multivariada (MVN). Este resultado se usa para proponer un método de imputación basado en este conocimiento nuevo.

Se evalúa el desempeño del método de imputación asumiendo que cada día se distribuye normal 24-variado. Y a partir de las propiedades de esta distribución se hallan las distribuciones de las horas faltantes, dadas las observadas, para, finalmente, generar un valor o vector de valores que se imputarán. La estimación de la matriz de varianzas y covarianzas se realiza por el método convencional de máxima verosimilitud y por el método de encogimiento (shrinkage) ([16], [17], [18]).

## II. MATERIALES Y MÉTODOS

### A. Datos faltantes e imputación

Sea  $X = (X_1, X_2, \dots, X_i, \dots, X_p)$  un vector aleatorio con función de probabilidad  $f_\theta$ , donde  $\theta$  es el vector de parámetros de interés. Y sea  $M = (M_1, M_2, \dots, M_i, \dots, M_p)$  un vector indicador de falta asociado a  $X$ , que toma el valor de 0 ó 1, para un  $X_i$

faltante u observado respectivamente. Si se define la falta de datos como variable aleatoria, se asume que hay una distribución de probabilidad que modela a  $M$  y una función o ecuación que describe la probabilidad de falta [17]. Así, la probabilidad de que  $M$  tome el valor de  $m = (m_1, \dots, m_p)$  dado que  $X$  toma el valor de  $x = (x_1, \dots, x_p)$  es  $g_\phi(m|x)$ . Donde  $\phi$  es un parámetro de perturbación y  $g_\phi$  es llamado mecanismo o proceso de datos faltantes. Así, al tener una realización de  $M$ ,  $\tilde{m} = \tilde{m}_1, \dots, \tilde{m}_p$  convenientemente se puede efectuar una partición en el vector aleatorio  $X$  de tal forma que  $\tilde{x} = (\tilde{x}_{(0)}, \tilde{x}_{(1)})$  donde los subíndices (0) y (1) son los indicadores de respuestas asociados a  $M$ . El objetivo es usar la información disponible  $\tilde{m}$  y  $\tilde{x}_{(1)}$  para obtener inferencias sobre  $\theta$  [19].

Ahora bien, al realizar el proceso de inferencia sobre  $\theta$ , se debe tener en cuenta la relación existente entre  $X$  y  $M$  ( $g_\phi$ ), para que las inferencias asociadas al vector de parámetros sean válidas. A continuación, se muestran las tres principales situaciones reconocidas en la literatura sobre esta relación:

#### 1) Pérdida completamente al azar: (MCAR)

Es el supuesto más general [20], en el que se afirma que el proceso generador es independiente de la variable de interés  $x$ , se define en (1).

$$g_\phi(m|x) = g_\phi(m) \quad (1)$$

Así, se puede demostrar (2)

$$f_\theta(x_{(1)}) = f_{\theta,\phi}(x_{(1)}|\tilde{m}) \quad (2)$$

Concluyendo que la función de la que se asume que proviene  $x$ , al ignorar el mecanismo de datos faltantes, es igual a la función de  $x$  condicionada a la variable  $m$ .

#### 2) Pérdida al azar: (MAR)

Este supuesto afirma que el proceso generador de faltas depende solo de los datos observados, como se describe en (3)

$$g_\phi(m|x) = g(m|x_{(1)}) \quad (3)$$

En este caso la falta de datos no depende de los valores perdidos  $x_{(0)}$ , solo depende de las variables observadas.

#### 3) Pérdida no aleatoria: (MNAR)

Dicho supuesto radica en que la presencia de un dato faltante en una variable  $Y$  depende tanto de sus valores observados como sus valores perdidos, que en términos de probabilidad se denota como aparece en (4)

$$g_\phi(m|x) = g_\phi(m|x_{(1)}, x_{(0)}) \quad (4)$$

Así pues, la probabilidad de que falten datos en  $Y$  depende tanto de los valores observados como de los valores perdidos. Este caso es el más desafortunado en la práctica ya que es un tipo de falta no ignorable. Se podría pensar que en estos casos la pérdida es determinística por un factor externo al

comportamiento de las observaciones haciendo que se pierda información para realizar imputaciones.

**B. Test de Little para probar MCAR**

Sea  $X = (X_1, \dots, X_i, \dots, X_p)$  el vector aleatorio de dimensión  $p$  y  $M = (M_1, \dots, M_i, \dots, M_p)$  como el vector indicador de falta. Entonces, se define a  $A$  como el número posible de distintos patrones de falta en  $X$ , donde los casos completamente observados cuentan como patrón. Además, al tener varias observaciones de  $X$ , se define  $X_a$  como el conjunto o matriz de vectores observados con patrones faltantes  $a$  ( $a = 1, \dots, A$ ) con  $r_a$  observaciones (filas) y  $p$  variables (columnas), cumpliéndose que  $\sum_a r_a = n$ . Por último, se define a  $D_a$  ( $p \times p_a$ ) (con  $p_a$  el número de variables observadas en el patrón  $a$ ), como la matriz **indicadora** de las variables que fueron observadas en el patrón  $a$ , que tiene una columna para cada variable presente que consta de  $p - 1$  ceros y un 1, correspondiente a la variable identificada. Consecuentemente, si se tienen  $n$  realizaciones  $X_k$   $k \in (1, \dots, n)$  iid normal, la pérdida de datos MCAR y estimadores por máxima verosimilitud tanto de  $\mu$  y  $\Sigma$   $\hat{\mu} = \bar{X}$  y  $\hat{\Sigma} = n * \frac{\hat{\Sigma}}{n-1}$  se procede a establecer el estadístico (5)

$$d_0^2 = \sum_a r_a (\bar{x}_{(1),a} - \hat{\mu}_{(1)a})^T \hat{\Sigma}_{(1),a}^{-1} (\bar{x}_{(1),a} - \hat{\mu}_{(1)a}) \quad (5)$$

Con  $\bar{x}_{(1),a} = r_a^{-1} \sum_{k \in X_a} x_{(1),k}$  siendo  $x_{(1),k}$  el vector de valores observados para un día  $k$  cualquiera ( $k = 1, \dots, n$ ),  $\hat{\mu}_{(1)a} = \hat{\mu} D_a$  y  $\hat{\Sigma}_{(1),a} = D_a^T \hat{\Sigma} D_a$ , medias y matrices de covarianzas de las variables observadas en el patrón  $a$ . Así pues, (2) se distribuye  $\chi^2$  con  $\sum_a p_a - p$  grados de libertad. Finalmente, Little demuestra que  $d_0^2$  es el estadístico de razón de verosimilitud para probar el modelo  $(x_{(1),k} | \tilde{m}_a) \sim N(\mu_{(1),a}, \Sigma_{(1),a})$  con  $k \in X_a$  contra el modelo alternativo modelo  $(x_{(1),k} | \tilde{m}_a) \sim N(v_{(1),a}, \Sigma_{(1),a})$  siendo  $v_{(1),a}$  el vector de medias de las variables observadas (diferentes de  $\mu$ ) que son distintos en cada patrón  $a$  [21].

**C. Imputación a partir de la distribución normal Multivariada**

La imputación de datos recoge a un conjunto de métodos para tratar bases de datos con datos faltantes. La idea principal de estos métodos consiste en sustituir los espacios de los datos faltantes por un dato estimado, para luego realizar análisis de datos [18]. Los tipos de imputación varían de acuerdo con las necesidades o alcances de la situación. A continuación, se presenta el método de imputación propuesto que se encuentra basado en el proceso generador (6).

$$X_j \sim N_{24}(\bar{X}_j, S_j); \quad j = 1, 2, \dots, 7 \quad (6)$$

Donde  $j$  se refiere a cada día de la semana comenzando por el lunes y terminando en el domingo.  $\bar{X}_j$  y  $S_j$  son los estimadores máximo verosímiles de los parámetros de la distribución normal multivariada de cada día.

De acuerdo con las propiedades de la normal multivariada, se realizan las particiones (7).

$$X_{j(p \times 1)} = \begin{bmatrix} X_{j1(q \times 1)} \\ X_{j2(p-q \times 1)} \end{bmatrix}; \quad \mu_{j(p \times 1)} = \begin{bmatrix} \mu_{j1(q \times 1)} \\ \mu_{j2(p-q \times 1)} \end{bmatrix}$$

$$Y$$

$$\Sigma_{j(p \times p)} = \begin{bmatrix} \Sigma_{j11(q \times q)} & \Sigma_{j12(q \times (p-q))} \\ \Sigma_{j21((p-q) \times q)} & \Sigma_{j22((p-q) \times (p-q))} \end{bmatrix} \quad (7)$$

En (7), se particiona el vector  $X_j$  en dos subvectores. Uno ( $X_{j1}$ ) se refiere a los datos que se requiere imputar y el otro ( $X_{j2}$ ) se refiere a las observaciones efectivamente obtenidas. No se pueden imputar más de 6 datos faltantes.

Por lo tanto, el método de imputación propuesto se basa en utilizar la distribución condicional de  $X_{j1} | X_{j2} = x_{j2}$  en un día  $k$  donde que contienen  $q_k$  datos faltantes ( $k = 1, \dots, n$ ). Así la idea es hallar la distribución normal  $q_k - variada$  que modela las horas faltantes en un día  $k$ , en el que se cuenta con información de las horas que observadas. Esta distribución se define en (8).

$$X_{j1} | X_{j2} = x_{j2} \sim \text{NMV}(\mu_{X_{j1} | X_{j2} = x_{j2}}, \Sigma_{X_{j1} | X_{j2} = x_{j2}}) \quad (8)$$

La media condicional y la matriz de covarianzas condicional se definen en (9) y (10).

$$\mu_{X_{j1} | X_{j2} = x_{j2}} = \bar{X}_{j1} + S_{j12} S_{j22}^{-1} (x_{j2} - \bar{X}_{j2}) \quad (9)$$

$$\Sigma_{X_{j1} | X_{j2} = x_{j2}} = S_{j11} - S_{j12} S_{j22}^{-1} S_{j21} \quad (10)$$

**D. Estimación de  $\mu$  y  $\Sigma$**

Para hallar los estimadores de  $\mu$  y  $\Sigma$  se usa el método de estimación por máxima verosimilitud teniendo el resultado (11).

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \quad (11)$$

$$S = \frac{n}{n-1} \hat{\Sigma}$$

Adicionalmente, para calcular  $\Sigma$ , se presenta un método alterno mediante el estimador de encogimiento (shrinkage), que se define en (12).

$$\Sigma^* = \lambda T + (1 - \lambda) S \quad (12)$$

Donde  $T$  es una matriz objetivo y representa un modelo reducido de la matriz de covarianzas en el que se estiman menos parámetros y  $\lambda \in [0, 1]$  representa la intensidad de contracción [17].

La elección de este estimador tiene que ver con el condicionamiento de la matriz, dado que en casos donde el tamaño de la muestra  $n$  no es suficientemente mayor a  $p$ , es decir para proporciones  $n/p$  cercanas o menores a 1, la matriz estimada puede perder algunas propiedades deseables [15], especialmente puede llegar a ser una matriz mal-condicionada. Cuando se presenta esta situación, se puede evaluar la matriz mediante número de condición  $k(\Sigma)$ , que para matrices de covarianza se define en (13).

$$k(\Sigma) = \frac{w_{max}}{w_{min}} \quad (13)$$

Donde  $w_{max}$  y  $w_{min}$  corresponden a los valores propios máximo y mínimo de  $\Sigma$ , respectivamente. Valores mayores de  $k(\Sigma)$  se asocian matrices  $\Sigma$  mal condicionadas.

#### E. Test de Mardia para probar el supuesto de Normalidad

Mardia [22] propone un test para probar normalidad multivariada a partir de las medidas de asimetría y curtosis multivariada. Estas medidas se desarrollan al extender aspectos de estudios para el estadístico t-student en las que se tienen en cuenta la simetría y la curtosis univariadas.

Sea  $\hat{\gamma}_1$  el coeficiente de simetría estimado de una muestra normal p-variada y  $\hat{\gamma}_2$  el coeficiente de curtosis estimado, se pretende validar el supuesto de normalidad p-variada de un vector aleatorio  $X = (X_1, \dots, X_p)$  a partir de una muestra de tamaño  $n$ , mediante los dos estadísticos (14) y (15).

$$\gamma_1 = \frac{1}{n^2} \sum_{b=1}^n \sum_{c=1}^n d_{bc}^3 \quad (14)$$

$$\gamma_2 = \frac{1}{n} \sum_{b=1}^n d_{bb}^2 \quad (15)$$

Donde  $d_{bc} = (x_b - \bar{x})^T S^{-1} (x_c - \bar{x})$  es la distancia de Mahalanobis al cuadrado entre la  $b$ -ésima y la  $c$ -ésima observación, y  $S^{-1}$  es la inversa de la matriz de varianzas y covarianzas estimada. Los estadísticos de prueba bajo la hipótesis de normalidad multivariante se distribuyen (16)

$$\frac{n}{6} \gamma_1 \sim \chi^2 \quad (16)$$

Con  $gl = \frac{p(p-1)(p+1)}{6}$  y la región de rechazo de la hipótesis de simetría normal p-variada está definida como  $R = (\chi^2 | \chi_{obs}^2 > \chi_{(\alpha, gl)}^2)$  siendo  $\chi_{(\alpha, gl)}^2$  el estadístico crítico asociado a los grados de libertad y una significancia  $\alpha$ . Ahora bien, Para la curtosis se tiene la distribución (17)

$$\gamma_2 \sim N\left(p(p+2), \frac{8p(p+2)}{n}\right) \quad (17)$$

Así la prueba para la curtosis se contrasta contra el valor crítico  $Z_\alpha$  asociado a una significancia  $\alpha$ .

#### F. Evaluación del desempeño

Para evaluar el método de imputación se hace uso de la raíz del error cuadrático medio (RMSE) (18) y el coeficiente de determinación ( $R^2$ ) (19).

$$RMSE = \sqrt{\frac{1}{H} \sum_{h=1}^H (\hat{x}_h - x_h)^2} \quad (18)$$

$$R^2 = \left( \frac{1}{H} \sum_{h=1}^H (\hat{x}_h - \bar{\hat{x}})(x_h - \bar{x}) \right)^2 \quad (19)$$

En las anteriores expresiones  $H$  es la cantidad total de horas que se imputaron,  $\hat{x}_h$  los valores de las horas imputadas y  $x_h$  las horas observadas,  $\bar{\hat{x}}$  y  $\bar{x}$  son la media de estos y  $\sigma_{\hat{x}}$ ,  $\sigma_x$  su desviación estándar. El RMSE da una noción de la distancia que hay entre los datos observados y los datos imputados, así que para que el método de imputación presente un buen desempeño esta distancia debería ser pequeña, por lo que el RMSE debe ser cercano 0. El  $R^2$  dice como es el ajuste de un modelo lineal entre los datos observados y los imputados, si este valor es cercano a 1 indica que los datos observados e imputados son equivalentes.

#### G. Generación de escenarios de simulación

En la referencia [15] proponen la generación de los escenarios usada en este trabajo de investigación, donde se tiene como objetivo generar una matriz de datos faltantes a partir de la matriz de datos completos que conserve los patrones de datos faltantes presentados en la matriz de datos originales, este método comienza haciendo un análisis de los datos faltantes en brechas horarias, que de acuerdo con el número de faltas se definen así:

Brecha de tamaño  $n$ . Una brecha de tamaño  $n \in \{1, 2, \dots, 6\}$ , consiste en  $n$  horas seguidas en las que hay datos faltantes pero antes y después de estas no se presentan horas con datos faltantes. El número de datos faltantes en un día no puede ser mayor del 25% de los datos posibles para el día, para que la información del día se pueda considerar válida. Por esta razón,  $n$  no puede ser mayor que 6.

Escenarios de simulación propuestos:

Contando con la información sobre las brechas horarias y haciendo uso de la distribución uniforme (0,1) se tiene el siguiente algoritmo para la generación de los escenarios:

- **Paso 1:** A partir de la matriz de días completos se escoge al azar un porcentaje  $T$  de estos días completos, con el propósito de contaminarlos con datos faltantes. Los días seleccionados serán contaminados usando las brechas ya

mencionadas. Los porcentajes de días a contaminar son  $T = \{20\%, 40\%, 60\%, 80\%, 100\%\}$ .

- **Paso 2:** Se toma el primer día elegido en el paso anterior y se genera un número de la distribución  $U \sim (0,1)$  y según el valor obtenido este día tendrá cierto tipo de brecha. Para esto se toma como referencia las frecuencias relativas acumuladas.
- **Paso 3:** Teniendo el tipo de brecha, se procede a elegir la cantidad de brechas que se van a generar dentro del día, de nuevo haciendo uso de la distribución  $U \sim (0,1)$ . Cabe mencionar que este paso sólo se realiza para las de brechas de tamaño 1, 2 y 3 ya que en el caso de las brechas de tamaño 4, 5 y 6 sólo es posible generar una brecha de cada cual debido a la restricción de que máximo pueden generarse 6 datos faltantes por día.
- **Paso 4:** Después de tener el tipo de brecha y su cantidad a generar en un día se eligen aleatoriamente las posiciones en las que estas se ubicarán.
- **Paso 5:** Se repiten los pasos 2, 3 y 4 para todos los días seleccionados en el paso 1.
- **Paso 6:** Se realiza el proceso de imputación propuesto a la base de datos contaminada.
- **Paso 7:** Teniendo la base de datos imputada, se procede a evaluar el método mediante el RMSE y el  $R^2$ .

Finalmente se repite el algoritmo 1.000 veces para cada porcentaje de datos faltantes expuestos en el paso 1. En cada iteración se guardan los valores del RMSE y el  $R^2$ . Entonces en total se tendrán 1.000 indicadores de cada uno por cada porcentaje  $T$ , de estos se reporta el promedio y el rango, con el fin de ver como es el comportamiento de estos indicadores para diferentes escenarios posibles.

### III. RESULTADOS

#### A. Datos

La información usada para ilustrar los métodos propuestos es tomada del Sistema de Vigilancia de la Calidad del Aire de Santiago de Cali (SVCASC), que cuenta con 9 estaciones distribuidas en las zonas urbana y rural, en las que se miden diferentes contaminantes atmosféricos incluido el  $PM_{2.5}$ . La base de datos consiste en mediciones horarias de  $PM_{2.5}$  del año 2018, tomadas en la estación Univalle [23]. El conjunto de datos cuenta con 365 filas y 24 columnas correspondientes a los días y horas respectivamente lo que implicaría un total de 8760 mediciones, en un contexto sin mediciones faltantes.

La estación Univalle del SVCASC está ubicada en la parte sur de la zona urbana de Cali (1018 msnm), cercana a vías principales como la Calle 5, la Carrera 100, la Calle 16, La Avenida Pasoancho (Calle 13) y la autopista Simón Bolívar. Esta zona se caracteriza por ser altamente transitada, ya que es camino hacia las principales instituciones de educación media y superior de la ciudad y hacia otras ciudades vecinas. Además,

está al lado de dos de los más grandes y concurridos centros comerciales de la ciudad, pero no hay industrias alrededor. De esta manera, se podría considerar a las fuentes móviles como las principales generadoras de partículas suspendidas en el aire en este sector de Cali.

#### B. Análisis exploratorio

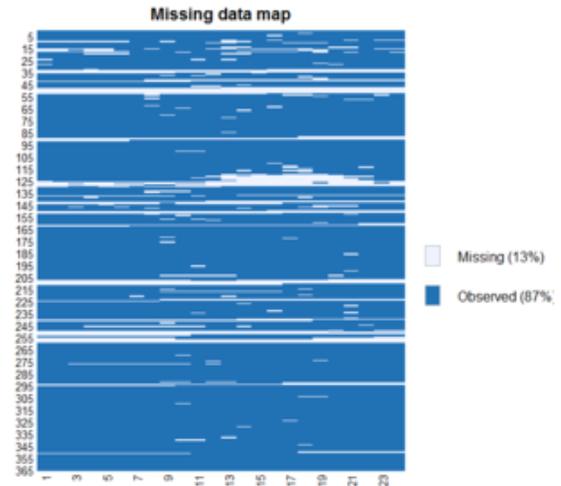


Fig. 1. Mapa de datos faltantes  
Fuente: Elaboración propia

La Fig. 1, representa el esquema de la base de datos mostrando de color blanco las horas donde faltan mediciones. Se observa un 13% de mediciones faltantes de  $PM_{2.5}$  de la estación de Univalle en el año 2019, un equivalente a 1.139 horas donde no se obtuvieron registros del contaminante. Las filas totalmente blancas representan días donde faltan todas las 24 mediciones, llamados días totalmente incompletos o simplemente días incompletos.

En la Tabla I se muestra el porcentaje de días completos para cada uno de los días, Como en la propuesta de imputación se asume que cada día tiene una distribución Normal 24-variada, se muestra la cantidad de filas que tendrá cada conjunto de datos completos, con el que se estimará la media  $\mu_j$  y la matriz de covarianzas  $\Sigma_j$  asociado al día  $j$  de la semana.

TABLA I  
DÍAS CON DATOS COMPLETOS

j	Día de la semana	Días Completos	Frecuencia Relativa
1	Lunes	26	0,50
2	Martes	34	0,65
3	Miércoles	33	0,63
4	Jueves	28	0,54
5	Viernes	32	0,62
6	Sábado	35	0,67
7	Domingo	32	0,62
	Total	220	

Fuente: Elaboración Propia

Los resultados de la Tabla I permiten anticipar posibles dificultades en el condicionamiento de la matriz de varianzas y covarianzas estimada, dado que  $p = 24$  y los valores de  $n$  fluctúan entre 26 y 35.

TABLA II  
CLASIFICACIÓN DE LOS DÍAS

Horas faltantes	Categoría	Frecuencia	Frecuencia Relativa	Frecuencia Relativa Acumulada
0	Completos	220	0,603	0,603
1-6	Parcial/Completos	84	0,230	0,833
7-23	Parcial/Incompletos	42	0,115	0,948
24	Incompletos	19	0,052	1
	Total	365	1	-

Fuente: Elaboración propia

De acuerdo con la Tabla II, se dispone de un 60,3% de días completos y de un 23% de los días parcialmente completos. Estos 84 días son los susceptibles de usar para efectos de imputación, de tal modo que, si se aplica la propuesta a la base de datos real, se imputarían los 84 días parcialmente completos, aumentando el número de días completos, ya sea por observación, o por imputación, del 60,3 al 83,3%.

La Tabla III muestra que los registros faltantes son más comunes los domingos (19%) y los lunes (15%). Este hecho podría anticipar una menor calidad de la imputación para estos días de la semana.

### C. Comportamiento del $PM_{2,5}$

Para la ilustrar el comportamiento del contaminante, se presentan dos boxplot correspondientes a las mediciones de concentración de  $PM_{2,5}$  para los días lunes (Fig. 2) y martes (Fig. 3).

TABLA III  
DATOS FALTANTES POR DÍA

Día de la semana	Cantidad (Horas)	Frecuencia Relativa
Lunes	190	0,15
Martes	111	0,09
Miércoles	139	0,11
Jueves	148	0,12
Viernes	141	0,11
Sábado	166	0,13
Domingo	243	0,19

Fuente: Elaboración propia

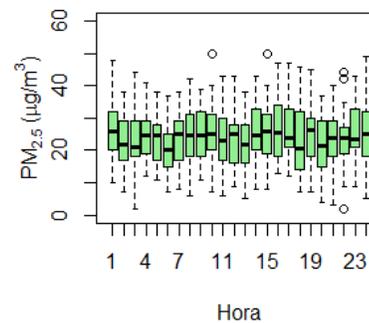


Fig. 2. Concentración horaria de  $PM_{2,5}$  para el día lunes.

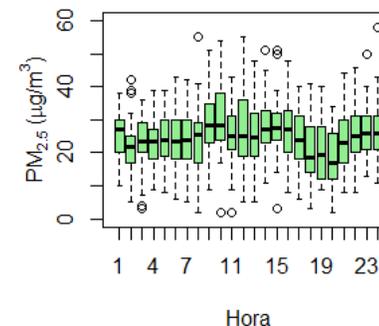


Fig. 3. Concentración horaria de  $PM_{2,5}$  para el día martes.

Las Fig. 2 y 3 permiten visualizar que las mediciones del contaminante incrementan alrededor de las 11 y de las 15 horas, posiblemente como resultado del tráfico vehicular alrededor de la estación Univalle. Al contrario, entre las 16 y 20 las concentraciones del  $PM_{2,5}$  disminuyen y se localizan en un rango menor.

### D. Estimación por encogimiento

Cuando las matrices de datos tienen el número de variables muy cercano al de individuos ( $p \approx n$ ), las matrices de covarianzas estimadas son en general mal condicionadas. Esta situación se presenta en este estudio, por lo que se presentan aquí la matriz objetivo  $T$  y el valor de  $\lambda$  que se utilizan como alternativa a la estimación por máxima verosimilitud.

Para la estimación de dichas matrices el parámetro de encogimiento es de  $\lambda = 0,2$ , por recomendación en la literatura [18]. Esta elección de  $\lambda$  es muy razonable, en la medida que da menor peso a la matriz objetivo. El cálculo de las matrices de covarianza estimadas se realiza con el lenguaje R [24] en RStudio [25] con la función “cov.shrink” ilustrada en [26], donde la matriz  $T$  se define en este caso de la forma (20).

$$T = \begin{cases} s_{ii} & si, i = j \\ 0 & si, i \neq j \end{cases}; i, j \in (1, \dots, 24) \quad (20)$$

La Tabla IV muestra los números de condición de las matrices de covarianzas estimadas por máxima verosimilitud ( $\hat{\Sigma}$ ) y por encogimiento ( $\hat{\Sigma}^*$ ). Conforme con lo esperado teóricamente, son mucho mayores para las estimaciones por máxima verosimilitud.

TABLA IV  
NÚMEROS DE CONDICIÓN PARA MATRICES DE COVARIANZA ESTIMADAS

Día	$k(\hat{\Sigma})$	$k(\hat{\Sigma}^*)$
Lunes	1.716,39	40,30
Martes	347,32	36,24
Miércoles	597,80	50,27
Jueves	2.104,46	41,50
Viernes	574,74	37,40
Sábado	864,68	37,13
Domingo	1.211,02	45,71

Fuente: Autores

### E. Pruebas de normalidad

En Tabla V se presentan los resultados de la prueba Mardia para probar el supuesto de normalidad 24-variada. Se busca contrastar la hipótesis de que las mediciones de cada día de la semana se asemejan a una distribución normal 24-variada a partir de las estimaciones de la simetría y curtosis multivariadas mostradas en la sección anterior.

TABLA V  
RESULTADOS TEST DE MARDIA

Día	Simetría		Curtosis		Cumple
	Estad.	Valor-p	Estad.	Valor-p	
Lunes	2.331,20	0,99	-3,33	0,000	SI/NO
Martes	2.559,89	0,90	-1,69	0,089	SI/SI
Miércoles	2.477,60	0,95	-2,32	0,020	SI/NO
Jueves	2.329,22	0,99	-3,24	0,010	SI/NO
Viernes	2.464,42	0,97	-2,32	0,019	SI/NO
Sábado	2.616,78	0,40	-1,27	0,200	SI/SI
Domingo	2.466,60	0,96	-2,43	0,014	SI/NO

Fuente: Autores

De acuerdo con los resultados de la Tabla V, en todos los días la prueba de simetría permite validar el supuesto de Normalidad Multivariada y en dos días la prueba de curtosis también permite mantener este supuesto. El test de Mardia se conduce por defecto con la estimación máximo verosímil de la matriz de covarianzas. En los datos de esta ilustración, estas matrices no son bien condicionadas, lo que podría afectar el desempeño de la prueba. Otras pruebas de Normalidad Multivariada, como la generalización de la prueba de Shapiro-Wilk [18] proceden de igual manera, por lo que los resultados de esta prueba también están afectados por esta realidad del mal-condicionamiento de estas matrices estimadas. Ante la ausencia de pruebas formales diseñadas para el escenario planteado por este conjunto de datos y ante los resultados del test de Mardia, sumada a la Normalidad de las distribuciones horarias, los resultados se basarán en una distribución Normal 24-variada para cada día de la semana.

### F. Prueba MCAR

A partir del supuesto de normalidad de cada día, en la prueba de Little para MCAR se busca contrastar las hipótesis (21)

$$H_0: f(y_j|m) \sim N(\boldsymbol{\mu}_j, \Sigma_j) \quad \text{vs.} \quad (21)$$

$$H_a: f(y_j|m) \sim N(\boldsymbol{v}_j, \Sigma_j)$$

La hipótesis alterna sugiere que la función de distribución de las concentraciones de  $PM_{2,5}$  en el aire, dado las faltas, surge de una distribución con media  $v_j \neq \mu_j$  en cada patrón  $j$  de faltas, probando que las faltas no son completamente aleatorias. En la Tabla VI se presentan los resultados.

TABLA VI  
RESULTADOS DEL TEST DE LITTLE PARA MCAR

	GL	Estadístico	Valor- p
Lunes	323	357,26	0,091
Martes	219	249,77	0,075
Miércoles	198	231,70	0,051
Jueves	284	302,14	0,219
Viernes	263	324,59	0,001
Sábado	153	194,31	0,013
Domingo	179	182,38	0,416

Fuente: Autores

Según los resultados de la Tabla VI, con un nivel de significancia del 5%, la condición MCAR no se cumpliría para viernes y sábado, mientras que para los días restantes no hubo evidencia para rechazar dicho supuesto. Es decir, las observaciones de estos días con registros faltantes se pueden considerar como submuestras aleatorias de la base completa, por lo que se podría realizar el método de imputación sin mayor complicación, porque gracias al cumplimiento del test, el mecanismo de faltas puede ser ignorado. Para la aplicación de la propuesta en los viernes y sábados se asume que la pérdida de datos es MAR. En este caso se repite la dificultad de que el test de Little se apoya en la estimación máximo verosímil de la matriz de covarianzas.

### G. Desempeño del método

En esta sección se presentan los resultados obtenidos del desempeño por el método de imputación propuesto para los diferentes escenarios de simulación planteados. Primero, en la Tabla VII se muestra el porcentaje de registros faltantes generados en las bases de datos para cada porcentaje de días contaminados.

TABLA VII  
PROPORCIÓN DE FALTAS GENERADAS PARA CADA PORCENTAJE DE DÍAS CONTAMINADOS

% días contaminados con faltantes	$\hat{p}$ (min - max)
20%	0,016 (0,011; 0,022)
40%	0,033 (0,024; 0,041)
60%	0,049 (0,041; 0,061)
80%	0,066 (0,051; 0,078)
100%	0,083 (0,071; 0,097)

Fuente: Autores

En la Tabla VII se muestra el promedio y el rango de las 1.000 simulaciones realizadas. Es evidente que a medida que aumenta el porcentaje de días contaminados el porcentaje de datos faltantes también aumenta, donde el máximo de faltas en una corrida fue de 9,7%, bajo un escenario del 100% de días contaminados.

En la Tabla VIII se muestran los resultados asociados al desempeño del método de imputación, evaluado con la raíz del error cuadrático medio (RMSE) y con el coeficiente de determinación, lo cual se aplicó para los diferentes niveles de días con horas faltantes, teniendo en cuenta las estimaciones de covarianza por máxima verosimilitud ( $\hat{\Sigma}$ ), y la estimación por encogimiento ( $\hat{\Sigma}^*$ ).

De acuerdo con los resultados de la Tabla VIII, los valores de  $\sqrt{RMSE}$  y  $R^2$  resultan similares bajo todos los escenarios incluyendo aquel con mayor porcentaje de faltas, por lo que se pueden interpretar las estimaciones dadas para el escenario con 100% de días contaminados con faltantes, generalizando el desempeño del método.

Se observa que la precisión del método es mejor al usar las estimaciones por máxima verosimilitud, comparada con las estimaciones por encogimiento. El error cuadrático medio toma un valor de 5,72 y 8,55 respectivamente para cada matriz, por lo que se afirma que en promedio las distancias entre los valores reales y los valores imputados se alejan en 5,72 unidades cuando se utiliza  $\Sigma$ ; este valor se incrementa a 8,55, más de 3 unidades, al usar  $\Sigma^*$ . El  $R^2$  promedio con  $\Sigma$  da como resultado 0,7, indicando que en promedio el 70% de la variabilidad total de las mediciones de  $PM_{2,5}$  se pueden explicar por las mediciones imputadas; en cambio, al usar  $\Sigma^*$  se nota que la explicación de la variabilidad total de las mediciones baja a 0,41, lo que indica una pérdida de explicación de la imputación del 30%.

TABLA VIII.  
DESEMPEÑO DEL MÉTODO DE IMPUTACIÓN

% días contaminados con faltantes	Matrices	$\sqrt{RMSE}$	$R^2$
20 %	$\hat{\Sigma}$	5,682 (3,910; 8,194)	0,698 (0,352; 0,875)
	$\hat{\Sigma}^*$	8,391 (6,224; 10,281)	0,434 (0,164; 0,635)
40%	$\hat{\Sigma}$	5,698 (4,374; 7,487)	0,698 (0,464; 0,829)
	$\hat{\Sigma}^*$	8,565 (7,201; 10,075)	0,411 (0,217; 0,596)
60%	$\hat{\Sigma}$	5,732 (4,479; 7,086)	0,699 (0,554; 0,808)
	$\hat{\Sigma}^*$	8,597 (7,407; 9,985)	0,410 (0,282; 0,568)
80%	$\hat{\Sigma}$	5,707 (4,618; 6,799)	0,700 (0,578; 0,804)
	$\hat{\Sigma}^*$	8,536 (7,510; 9,320)	0,417 (0,325; 0,526)
100%	$\hat{\Sigma}$	5,718 (4,761; 6,673)	0,700 (0,581; 0,795)
	$\hat{\Sigma}^*$	8,556 (7,523; 9,996)	0,411 (0,281; 0,507)

Fuente: Autores

Con las mediciones horarias de  $PM_{2,5}$  disponibles, más las mediciones imputadas, se construyen los boxplot que aparecen en las Fig. 4 y 5 para lunes y martes.

Estos resultados permiten ver cómo las imputaciones no afectan el comportamiento observado de las concentraciones del contaminante para ambos días. Los resultados para los demás días son similares, aunque no se presentan por razones de espacio.

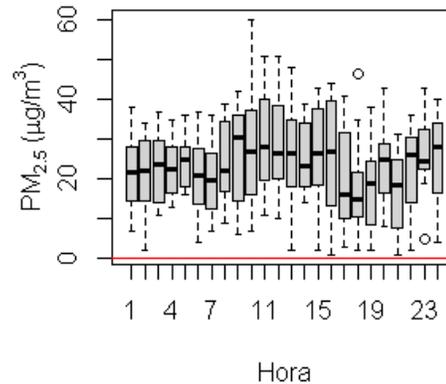


Fig 4. Concentraciones horarias observadas e imputadas de  $PM_{2,5}$  para el lunes

#### IV. CONCLUSIONES

A partir de los datos en la ciudad de Cali en 2018, estación Univalle, los resultados muestran que las mediciones de partículas finas suspendidas en el aire ( $PM_{2,5}$ ) efectivamente cuentan con datos faltantes, tal como se anticipa en la literatura. Para el 2018 se identificó un 13% de faltantes horarias y un total de 19 días en los que no hubo ninguna medición. Ahora bien, dados los hallazgos en los últimos años de la relación del material particulado con la morbilidad en enfermedades cardiovasculares es importante hacer un seguimiento de calidad a la contaminación del ambiente en la ciudad de Cali. Por lo tanto, se concluye, en primera medida, que se debe persistir en los esfuerzos por reducir esta generación de datos faltantes para conservar las propiedades de los estimadores utilizados. Y, en segundo lugar, que resulta importante el estudio permanente del comportamiento del contaminante, para continuar investigando métodos de imputación que contribuyan a reducir las mediciones faltantes, con el fin de conservar la calidad de los estimadores.

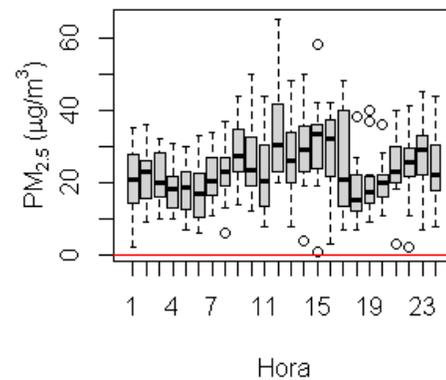


Fig 5. Concentraciones horarias observadas e imputadas de  $PM_{2,5}$  para el martes

Los escenarios de imputación propuestos buscan recoger las mismas características del patrón de faltas de la matriz original,

teniendo presente que estos guardan propiedades MCAR, ya que las faltas se generan sin condiciones que impliquen un tipo relación entre las faltas y los valores observados. Bajo los escenarios de simulación propuestos, el método de imputación obtuvo buen desempeño, similar bajo todos los escenarios de simulación, incluso en los escenarios donde se contaminó el 100% de los días.

Debido al reducido tamaño muestral para estimar las matrices de covarianza con un tamaño relativamente grande de variables, el estimador de encogimiento resultó ser una herramienta que permitió generar mejores imputaciones, situación en la que se prefirió perder la precisión del método con el fin de ganar imputaciones con resultados lógicos. Podría ser de interés estudiar si otras matrices objetivo mejoran el desempeño del método.

Un estudio que se deriva de los resultados es la evaluación de los test de Normalidad Multivariante y de MCAR, en las condiciones que se evidenciaron en este trabajo, con  $p \approx n$ , dadas las dificultades con la estimación de las matrices de covarianzas.

Como se mencionó en la introducción, con los resultados obtenidos se puede proponer que las observaciones horarias de  $PM_{2,5}$  de cada día sean modeladas usando un proceso gaussiano ya que este se define como una generalización de la distribución normal multivariada.

La ausencia de pruebas específicas de Normalidad multivariante para los casos de con  $p \approx n$  se convierten en una potencial debilidad del procedimiento propuesto. Además, el hecho de que el estimador máximo verosímil de la matriz de covarianzas no sea admisible [27], implica que necesariamente se debe pensar en alternativas para el contraste de hipótesis sobre Normalidad multivariante, que puedan mejorar las condiciones para el uso de la propuesta presentada en este trabajo.

## REFERENCIAS

- [1] OMS, “9 de cada 10 personas en todo el mundo respiran aire contaminado, pero más países están tomando medidas”, Organización Mundial de la Salud. Ginebra Suiza, Comunicado de prensa. Disponible: <https://www.who.int/es/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action>, consultada en marzo 11, 2023
- [2] IQAir, “PM<sub>2,5</sub>”, Actualización más reciente septiembre 22, 2015. Disponible: <https://www.iqair.com/newsroom/pm2-5>.
- [3] Cao, Junji and Chow, Judith and Watson, John and Lee, Shuncheng, “A brief history of PM<sub>2.5</sub> and its adverse effects”. *Aerosol and Air Quality Research*, Ene. 2013. DOI:10.4209/aaqr.2012.11.0302
- [4] Observatorio Nacional de Salud, “Carga de enfermedad ambiental en Colombia”. *Instituto Nacional de Salud (INS)*, pág. 96 Bogotá D.C. Nov. 2018. Disponible: <https://www.ins.gov.co/Noticias/Paginas/Informe-Carga-de-Enfermedad-Ambiental-en-Colombia.aspx>
- [5] M. E. Quinteros, S. Lu, C. Blazquez, J. P. Cárdenas-R Ossa, X., Delgado-Saborit, J. M., Harrison, R. M. and Ruiz-Rudolph, P., “Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile”, *Atmospheric Environment* 200, pp. 40-49. 2019. DOI: 10.1016/j.atmosenv.2018.11.053
- [6] Pope III, C. A., J. B., Anderson, J. L., Cannon, J. B., Hales, N. M., Meredith, K. G., Le, V. and Horne, B. D., “Short-Term” exposure to fine particulate matter air pollution is preferentially associated with the risk of ST-Segment elevation acute coronary events”. *Journal of the American heart association*. 2015. DOI: 10.1161/JAHA.115.002506.
- [7] Beyea, J., Stellman, S. D., Teitelbaum, S., Mordukovich, I. and Gammon, M. D. “Imputation method for lifetime exposure assessment in air pollution epidemiologic studies”, *Environmental Health*. 2013. DOI: 10.1186/1476-069X-12-62
- [8] M. Lee, P. Koutrakis, B. Coull, I. Kloog, and J. Schwartz., “Acute effect of fine particulate matter on mortality in three Southeastern states from 2007-2011”, *Journal of exposure science & environmental epidemiology*, pp 173-179. 2015. DOI: 10.1038/jes.2015.47
- [9] S. M. Taghavi-Shahri, A. Fassó, B. Mahaki and H. Amini, “Concurrent spatiotemporal daily land use regression modeling and missing data imputation of fine particulate matter using distributed space time expectation maximization”, *bioRxiv*. DOI: 10.1101/354852
- [10] J. Céspedes., J. Cuero and F. Hernández “Metodología para seguir las concentraciones de aerosoles atmosféricos usando técnicas de teledetección”, *Universidad del Valle*, Colombia. Sep. 2015.
- [11] L. C. Chien, Y. A. Chen and H. L. Yu, “Lagged Influence of fine particulate matter and geographic disparities on clinic visits for children’s asthma in Taiwan”. *International journal of environmental research and public health*. Abr. 2018. DOI: 10.3390/ijerph15040829
- [12] D. Allison, “Quantitative Applications in the Social Sciences: Missing data”. *Univ. of Pennsylvania, Pennsylvania P, USA*, 2002. DOI: 10.4135/9781412985079
- [13] N. A. Zakira, and M. N. Noor, “Imputation methods for filling missing data in urban air pollution data for Malaysia”. *Urbanism, Arhitectură. Construcții, Malaysia*. Vol 9, No. 2, 2018.
- [14] A. Caicedo and C. Jiménez, “Imputación basada en análisis de datos funcionales de observaciones faltantes de contaminación atmosférica por partículas finas suspendidas en el aire (PM<sub>2,5</sub>)”. *Universidad del Valle*, Colombia. 2016.
- [15] A. Otero, and M. Presiga. “Evaluación de un método de imputación basado en el Análisis de Datos Funcionales para los registros de PM<sub>2.5</sub> en la ciudad de Cali”. Trabajo de grado en Estadística, *Universidad del Valle*, Colombia. Dic. 2019.
- [16] G. G. Fernando. “Estimación de matrices de covarianzas: nuevas perspectivas”, Universidad Nacional de Educación a Distancia, España, 2014. Disponible: <http://espacio.uned.es/fez/eserv/bibliuned:masterMatavanz-Fgodino/Documento.pdf>
- [17] J. Schäfer and K. Strimmer. “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics”. *Statistical applications in genetics and molecular biology*, vol. 4, Feb. 2005. DOI: 10.2202/1544-6115.1175
- [18] J. Villaseñor and E. Gonzales, “A Generalization of Shapiro–Wilk’s Test for Multivariate Normality”. *Communication in Statistics - Theory and Methods*, 2009. DOI: 10.1080/03610920802474465
- [19] C. K., Enders “Applied Missing Data Analysis”. *Univ. of Pennsylvania*, New York, NY, USA, 2010. Disponible: <http://hsta559s12.pbworks.com/w/file/attach/52112520/enders.applied>
- [20] Rubin and B. Donald, “Inference and missing data”. *Biometrika* vol. 63 pp. 581-592. Oxford University Press, 1976. DOI: 10.2307/2335739
- [21] Little and J. A. Roderick, “A Test of Missing Completely at Random for Multivariate Data with Missing Values”, *Journal of the American Statistical Association*, vol. 83, pp. 1198 - 1202. Dic. 1988. DOI: 10.1080/01621459.1988.10478722
- [22] Mardia and V. Kanti, “Measures of multivariate skewness and kurtosis with applications”, *Biometrika* vol. 57, no. 3, pp. 519-530. Dic. 1, 1970. DOI: 10.1093/biomet/57.3.519
- [23] DAGMA, “Sistema de Vigilancia de Calidad del Aire de Cali - SVCAC” Cali, Colombia, acceso: Julio 2020.
- [24] R Core Team, “R: A Language and Environment for Statistical Computing” Viena, Austria. 2020 URL: <https://www.R-project.org>
- [25] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL: <http://www.rstudio.com/>.
- [26] J. Schäfer and R. Opgen-Rhein and V. Zuber and M. Ahdesmaki and P.D. Silva and K. Strimmer (Maintainer). “Package corpcor”. *R Package Versión 1.6.9*. Ene. 4, 2017. DOI: 10.2202/1544-6115.1175

- [27] H. Tsukuma and T. Kubokawa, "Shrinkage Estimation for Mean and Covariance Matrices". Springer, 2020. DOI: 10.1007/978-981-15-1596-5



**Esteban Arroyave López** nació en Cali, Colombia y obtuvo su título de Estadístico en la Universidad del Valle, ubicada en esta misma ciudad, en el año 2021. Actualmente trabaja como contratista en el proyecto Big Data del departamento de las TIC, en la Alcaldía de la de Santiago de Cali,

brindando asesorías técnicas transversales a la entidad para guiar la formulación e implementación de casos de uso de analítica avanzada mediante computo en la nube. Entre sus intereses se encuentra la gestión de proyectos de analítica y los modelos de aprendizaje automático.

ORCID: <https://orcid.org/0000-0001-6844-0828>



**Alejandro Villarreal Monsalve** nació en la ciudad de Cali, es Estadístico de la Universidad del Valle, graduado en el año 2021. Actualmente desempeña el cargo de Analista de Analítica de datos en Seguridad Atlas LTDA, su rol principal consiste en apoyar el desarrollo de soluciones analíticas en la empresa, empleando

metodologías de estadística, ciencia de datos y computación en la nube. Sus intereses son ingeniería de datos y la integración de modelos de analítica en aplicaciones de software para el usuario final.

ORCID: <https://orcid.org/0000-0003-1304-2958>



**Javier Olaya Ochoa** recibió su título de pregrado en Estadística de la Universidad del Valle, Cali, Colombia, en 1986. Sus títulos de Maestría en Ciencias Matemáticas, en 1997, y el de PhD en Management Science, en 2000, de la Universidad de Clemson, SC, USA.

Actualmente es profesor titular en la Escuela de Estadística de la Universidad del Valle, Cali, Colombia. Sus intereses de investigación incluyen los métodos de suavización y regresión no paramétrica, el análisis de datos funcionales y las aplicaciones de la Estadística en problemas ambientales, especialmente de la calidad del aire. Correo electrónico:

ORCID: <https://orcid.org/000-0001-7014-2782>