Pontificia Universidad JAVERIANA Bogotá

Univ. Sci. 25 (1): 1-16, 2020.
doi: 10.11144/Javeriana.SC25-1.faov

ORIGINAL ARTICLE

Functional analysis of variance of air pollution caused by fine particles

Javier Olaya Ochoa[1,*], **Diana Paola Ovalle Muñoz**[1], **Cristhian Leonardo Urbano León**[1]

Edited by
Juan Carlos Salcedo-Reyes
(salcedo.juan@javeriana.edu.co)

1. Escuela de Estadística, Facultad de Ingeniería, Universidad del Valle, Cali, Colombia.

* javier.olaya@correounivalle.edu.co

Received: 01-06-2018
Accepted: 29-08-2019
Published on line: 28-01-2020

Citation: Olaya Ochoa J, Ovalle Muñoz DP, Urbano León CL. Functional analysis of variance of air pollution caused by fine particles, *Universitas Scientiarum*, 25 (1): 1-16, 2020. doi: 10.11144/Javeriana.SC25-1.faov

Funding:
N.A.

Electronic supplementary material:
N.A.

Abstract

Environmental pollution is harmful to human health, as it can lead to chronic respiratory diseases. In particular, fine particles suspended in the air ($PM_{2.5}$) count among the most aggressive air pollutants. $PM_{2.5}$ levels vary depending on local conditions. The goal of this work was to compare year-round airborne $PM_{2.5}$ readings from three air quality surveillance stations in Cali (Colombia) to determine whether these show significant spatial and temporal variation. We subjected the obtained $PM_{2.5}$ dataset to a functional analysis of variance. We observed that $PM_{2.5}$ levels vary significantly among the three measurement sites on a temporal scale. Whereas in the morning hours $PM_{2.5}$ levels among the three sites differed most, in the afternoon and evening hours, the corresponding $PM_{2.5}$ levels were not significantly different.

Keywords: Airborne particles; environmental pollution; functional data.

Introduction

Particulate matter (PM) constitutes the main pollutant in the air and is closely related to the emergence of diseases (Bell *et al*., 2007). PM is usually classified based on aerodynamical diameter, fine particles in the air have aerodynamic diameters up to 2.5 μm and are known as $PM_{2.5}$ (Laden *et al*., 2000; Febrero-Bande *et al*., 2007; Al-Hamdan *et al*., 2009. $PM_{2.5}$ are associated to chronic respiratory illnesses such as asthma and lung cancer, as well as to cerebrovascular diseases, among others.

To ensure air quality in urban areas, public and private organizations undertake the monitoring and controlling the amount of PM in the air. In the city of Cali, Colombia, the environmental authority in charge is the Administrative Department of Environmental Management (DAGMA). DAGMA runs the city's Air Quality Monitoring System (SVCASC). The SVCASC consists of nine monitoring stations surveilling air quality variables. However, out of the nine stations, three of them asses $PM_{2.5}$. These three stations are named Compartir (denoted CO on this paper), Base Aérea (BA), and Universidad

del Valle (UV). Station CO is located at the most east-side residential zone of the city. It has a heavy impact from mobile sources, especially early in the morning and early in the evening. Station BA is in the city's northeast and it is located in the middle of a small industrial area, it is also close to an air force base. Station UV is in the city's southside, in the middle of the Universidad del Valle main campus, with a moderate impact from mobile sources. It is encircled by a highly residential and commercial neighborhood.

We analyzed the $PM_{2.5}$ air pollution data from the stations CO, BA, and UV. The comparison may help the local environmental authority decide on whether to keep recording measurements at these three sites. This approach provides ground to suggest the redesigning of the local surveillance network. In addition, our results may be used as a guide for the imputation of some missing points at one or more places. In this $PM_{2.5}$ air pollution data study, we used a functional extension of a classical Analysis of Variance (ANOVA), that compares the variances within each station related to the variance among stations and that bases decision on an F-test. This extension is known as FANOVA (Ramsay & Silverman, 2005), its corresponding p-value is a continuous curve, allowing the user to identify those hours at which differences are statistically significant.

We employed discrete data, namely hourly measurements, which are assumed to come from an unknown continuous function. The response is a daily contamination curve. The set of indicator variables representing the stations, are used as predictors. Some recent work on this paper's main topic can be seen in Górecki & Smaga (2017), Górecki & Smaga (2018), Ruiz-Medina (2016), Estévez-Pérez & Vilar (2013) and Zhang (2013).

Materials and Methods

Particulate matter data

The available data are the $PM_{2.5}$ hourly averages at three stations in 2015. For each day without missing points, there would be 24 observations at each station. Thus, the fine particulate matter levels are discretely measured in time. According to Ramsay & Silverman (2005), a functional data analysis (FDA) is appropriate when measurements satisfy this condition, enabling the adequate representation of the data such that a detailed analysis can be performed. FDA is advantageous when considerable amounts of information need to be analyzed (Ramsay & Silverman, 2005).

The SVCASC takes air contaminant readings every 10 seconds; however, it only reports the $PM_{2.5}$ average per hour in $\mu g/m^3$. Consequently, we had expected to retrieve 24 observations per day, during 365 days for each station. However we obtained data corresponding to 164 days for station CO, 82 days for station BA, and 167 days for station UV. Missing data is a common problem on these measurement systems.

Functional Data Analysis

Ferraty & Vieu (2006) formally define a functional random variable: "A random variable χ is called a functional variable if it takes values in an infinite dimensional space (or functional space). An observation x of χ is called a functional datum". In this sense, FDA inherits the descriptive and inferential statistics procedures extended from the scalar case to functions. To proceed with the FDA, the first step is to convert the discrete measurements into a smooth curve through smoothing techniques using linear combinations of a collection of functions. In Ramsay & Silverman (2005), a basis function system is defined as a set of known functions f_j, with $j = 1, 2, ..., k$, that are linearly independent and belongs to a functional space. Because the goal is to perform a comparison, it is necessary to work in a functional space that is defined over the field of real numbers and that has a norm and inner product, allowing for the notion of an orthogonal and orthonormal basis as an extension of the concept of linear independence.

On the other hand, Hastie *et al.* (2009) and James *et al.* (2013) mention that the most used basis function systems for the construction of smooth curves are Fourier series and B-splines. The former seem to be more appropriate for data with periodic or cyclic behaviors, whereas B-splines are easily adapted to behaviors with local changes, so their use extends to different fields. Ramsay & Silverman (2005) defines a spline as a polynomial function whose flexibility allows it to easily adapt to the data's behavior. The functions of the Fourier series and B-splines form a set of functions $\{f_j\}_{(j=1)}^{\infty}$. However, a functional data analysis is conducted on a finite subspace denoted by $\{f_j\}_{(j=1)}^{k}$. In the B-splines case, each function is defined by an order m and a sequence of nodes τ, and any spline function can be expressed as a linear combination of the basis functions.

For the construction of the functional data, it is important to confirm that the measurements have been taken discretely. They are denoted z_h, with $h = 1, 2, ..., n$, where n is the total number of measurements. Subsequently, the model of equation (1) is fitted, where t_h is the $h - t\,h$ time at which the measurement z_h was taken, c_j are coefficients to be estimated, and ε_h is a random error.

$$\hat{Z}_h = c_1 f_1(t_h) + c_2 f_2(t_h) + c_3 f_3(t_h) + ... + c_k f_k(t_h) + \varepsilon_h. \quad (1)$$

Because there are n observations in total, the smoothing can be represented in a matrix form as in equation (2) where,

$$F = \begin{pmatrix} f_1(t_1) & . & . & . & f_k(t_1) \\ f_1(t_2) & . & . & . & f_k(t_2) \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ f_1(t_n) & . & . & . & f_k(t_n) \end{pmatrix}_{n \times k}$$

and

$$C = \begin{pmatrix} c_1 \\ c_2 \\ . \\ . \\ . \\ c_k \end{pmatrix}_{k \times 1.}$$

$$y = FC + \varepsilon. \tag{2}$$

It is important to identify the appropriate number k of functions in the basis, which, according to Febrero-Bande & Oviedo de la Fuente (2012), can be performed using the Generalized Cross Validation (GCV) methodology. Thus, the number k that minimizes equation (3) will be the optimum number of basis functions that should be chosen.

$$GCV(k) = \frac{n^{-1} SSE}{[1 - tr(s_k) n^{-1}]}, \tag{3}$$

In equation (3), $tr(s_k)$ is the trace of the smoothing matrix s_k, and SSE is the error sum of squares defined as:

$$SSE = \frac{\sum_{h=1}^{n} (z_h - \hat{z}_h)^2}{n}.$$

Smoothing of functional data

After defining the system of basis functions to be used, it is important to identify the estimation method for finding the c_j coefficients of the model in equation (2), which are implicit in the vector C. To do so, different methods are proposed, including Ordinary Least Squares (OLS), Weighted Least Squares (WLS), and Penalized Spline Smoothing (P-Spline).

The OLS method seeks to minimize the sum of squares of the distances between the observed value and the fitted value, considering that the variance of the errors is constant; however, it is preferable to use the WLS method because it allows involving the variance and covariance matrix in the model when the errors exhibit autocorrelation structures. This matrix is generally denoted as Σ_ε (Ramsay & Silverman, 2005) and in this case, the estimation of the coefficient vector C is given by equation (4), which is represented in a matrix form.

$$\hat{C} = (F'WF)^{-1}F'Wy. \tag{4}$$

In equation (4), F is a matrix of size $n \times k$ that contains the values of the basis functions evaluated at time t_h, which is $f_j(t_h)$. W is a weight (or weighting) matrix taken as $(\Sigma_\varepsilon)^{-1}$ of size $n \times n$. Finally, the constructed functional datum is given by equation (5).

$$\hat{y} = d = F\hat{C} = F(F'WF)^{-1}F'Wy. \tag{5}$$

However, the P-Spline method adds to the OLS method (which considers $W = I$) a penalty for lack of smoothness given by the parameter λ, which can be estimated using GCV by only varying λ. In this case, the estimated coefficients are the solutions to the minimization of equation (6).

$$n^{-1}\sum_{h=1}^{n}(Z_h - \hat{Z}_h)^2 + \lambda\int_0^1 \hat{y}''(t)^2 dt. \tag{6}$$

In practice, there are N functional data. Once the smoothed curves have been obtained, the following step is the calculation of their functional mean and variance, which are defined in equations (7) and (8), respectively, where N is the total number of curves in the sample.

$$\bar{y}(t) = \frac{\sum_{i=1}^{N}\hat{y}_i(t)}{N}. \tag{7}$$

$$var[y(t)] = \frac{\sum_{i=1}^{N}[\hat{y}_i(t) - \bar{y}(t)]^2}{N-1}. \tag{8}$$

Modeling in FDA

Sometimes, it is of interest to identify whether the variation of a functional variable can be explained by a model based on other independent or regression variables. To do so, it is appropriate to fit functional regression models. In this

case, the model considers a functional response and indicator covariates. The model is very useful in studies involving the identification of characteristics of the response variable according to variables represented in factors. Ramsay & Silverman (2005) illustrate the case with functional temperature modeling, evaluating the effect of the geographic zone in which the weather station is located. In this case, it is necessary to perform a FANOVA because the response variable is functional.

The model is formally given by equation (9), where μ is the global mean of $\hat{y}$ without considering the effect of the treatment or of the group g (in our case, stations CO, BA and UV), G is the number of groups, α_g is the specific effect of group g on the response variable, and ε_i is the unexplained variation in individual i, $i = 1, 2, ..., N$.

$$\hat{d}_i(t) = \mu(t) + \alpha_g(t) + \varepsilon_i(t). \tag{9}$$

Thus, $\hat{d}_i$ is $\hat{\hat{y}}_i$ and H is defined as the design matrix of size $N \times (G+1)$; if $\beta_1 = \mu$, and $\beta_{(G+1)} = \alpha_G$, then there is the functional parameter vector $\beta = (\mu, \alpha_1, ..., \alpha_G)'$.

In this way, the model of equation (9) can be represented as in equation (10). On the other hand, the matrix representation of model (10) is $D = H\beta + \varepsilon$; therefore, using the general linear model theory, the estimator vector $\hat{\beta}$ is obtained via OLS as in equation 11.

$$\hat{d}_i(t) = \sum_{j=1}^{G+1} H_{ij}\beta_j(t) + \varepsilon_i(t), i = 1, 2, ..., N. \tag{10}$$

$$\hat{\beta} = (H'H)^{-1}H'D. \tag{11}$$

In the functional field, the analysis of variance table looks exactly the same as a scalar data table; its main difference lies in that there are no scalar sums of squares but curves as seen in the **Table 1**. Owing to this analysis, the hypothesis testing is performed, which allows identifying whether some of the parameters associated with one of the groups are statistically significant. The statistical hypotheses are:

$$H_0 : \beta_2 = \beta_3 = ... = \beta_{G+1} \; vs \; H_1 : \beta_j - \beta_i \neq 0,$$

for some $i, j = 2, ..., G+1, i \neq j$.

Table 1. Pointwise FANOVA.

Source of Variation	Sum of Squares	Degrees of freedom(df)	Means Square	F_0
Regression	$SSR(t)=\sum_{i=1}^{N}[\hat{d}_i(t)-\bar{y}(t)]^2$	$df_R=G-1$	$MSR(t)=\frac{[SST(t)-SSE(t)]}{df_R}$	$F(t)=\frac{MSR(t)}{MSE(t)}$
Residuals	$SSE(t)=\sum_{i=1}^{N}[d_i(t)-\hat{d}_i(t)]^2$	$df_E=N-G$	$MSE(t)=\frac{SSE(t)}{df_E}$	
Total	$SST(t)=\sum_{i=1}^{N}[d_i(t)-\bar{y}(t)]^2$	$df_T=N-1$		

Test statistics $F(t)$ (Table 1) is the F—statistics evaluated pointwise, as proposed by Ramsay & Silverman (2005). Thus, we compute the pointwise p-value as follows:

$$p(F_{G-1,N-G}>F(t)).$$

The procedures to construct the functional data needed for the FANOVA analysis are carried out using The R Project packages *fda* (Ramsay *et al*., 2018) and *fda.usc* (Febrero-Bande & Oviedo de la Fuente, 2012).

Results and Discussion

To convert 24 discrete hourly measurements from one day into a smooth curve, we selected the B-spline approach, since we do not have evidence of any kind of periodicity in the data, and have empirical evidence of a lack of periodicity for some days of the week. We also chose the cubic polynomials approach, because of its flexibility and adaptability, given the high variability of the response. B-splines have been chosen by some other authors (Al-Hamdan *et al*., 2009) for fine particulate matter analysis based on FDA. Thus, the daily curves of the available complete days for the entire year were constructed. For this, we picked the best number k of third-degree B-splines to form a smooth-curves generating set. That selection is based on the GCV criterion, which minimizes the mean quadratic error of the smooth curve estimator. Similarly, the value of the penalization parameter λ is chosen, in this case, the optimum values are $k=20$ and $\lambda=1$.

In addition, the weekday variable was introduced in the analysis such that seven separated models are constructed, one for each day of the week, seeking to refine the analysis for determining whether the curves of the three stations are significantly different. The number of complete days (those with 24 valid observations) is not the same at each station because of randomness of missing data points. Similarly, neither we have the same number of complete days. It means that we do not have the same number of curves per station, nor per day. Thus, the optimum values of the number of basis functions k and the smoothing parameter λ (modes are chosen in both cases) are $k=15$ and $\lambda=1$.

Analysis of results with smoothing parameters without specifying day of the week

Fig. 1 shows the observed data points (Fig. 1A) together with the smoothed data (Fig. 1B). The hourly averages per station (Fig. 1C) and the functional mean (Fig. 1D) are also depicted.

The current Colombian air quality regulation (Resolución 2254 de 2017, Ministerio del Medio Ambiente y Desarrollo Territorial) dictates that the maximum hourly average allowed for $PM_{2.5}$ is 37 $\mu g/m^3$. We compared the observed hourly averages against the standard using datapoints (Fig. 1C) as well as their functional means (Fig. 1D). We noted that the observed hourly averages never exceeded the established upper limit at any of the three stations.

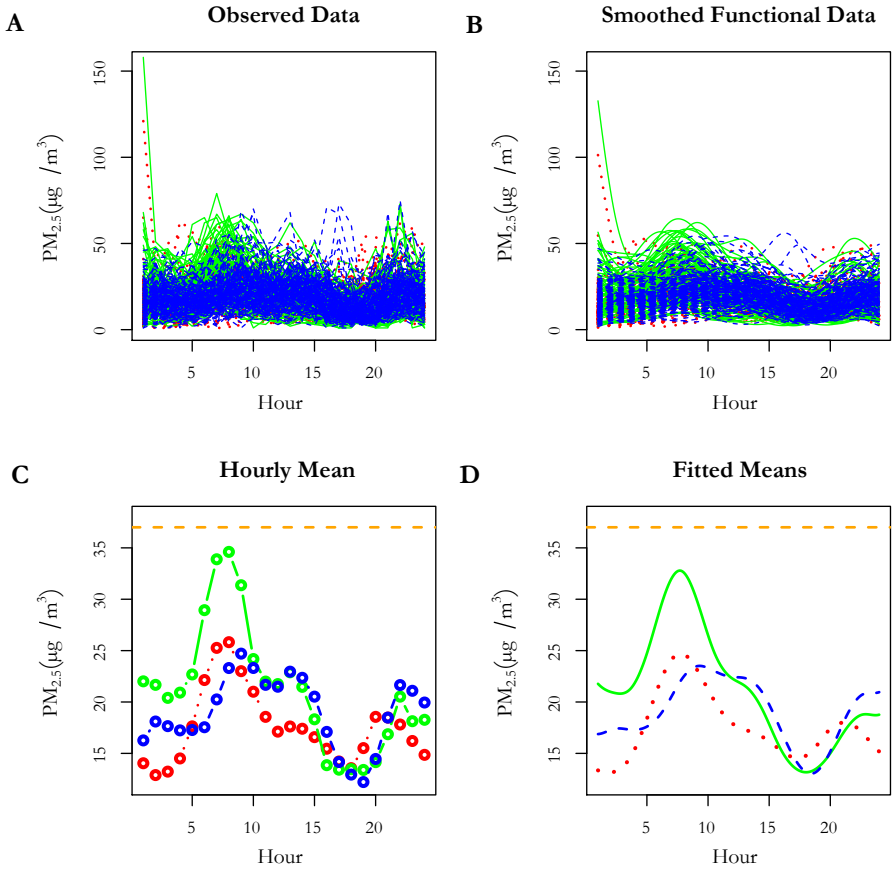

Figure 1. Observed and Functional Particulate Matter ($PM_{2.5}$) data are shown in panels A and B, respectively, along with hourly averages and functional means (Panels C and D). Green color is used for Station BA, red for Station CO, and blue for Station UV. Orange dotted lines on panels C and D indicate the Colombian standard for $PM_{2.5}$.

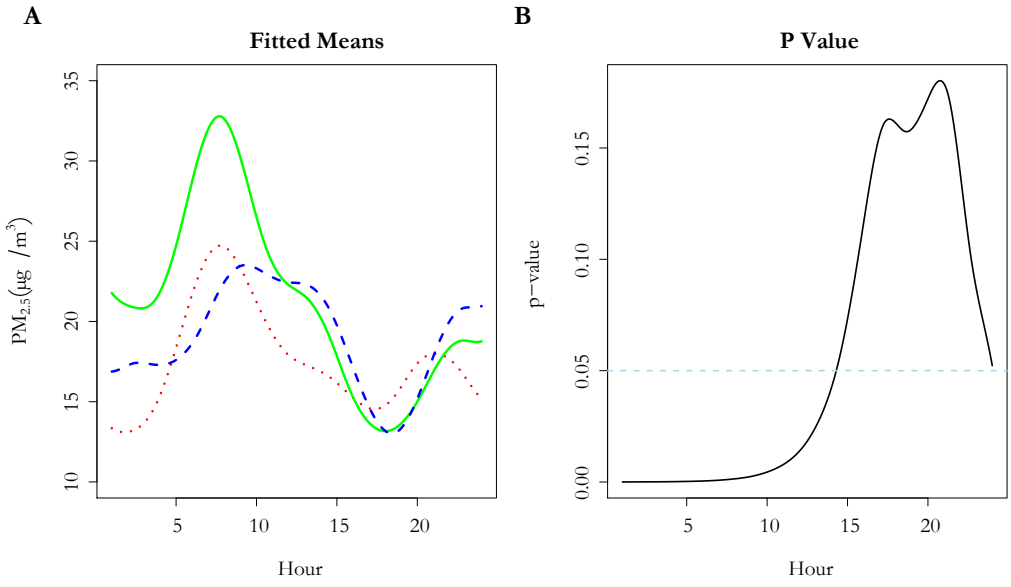

Figure 2. Results for the functional analysis of variance for the entire year. The green curve (solid line) in panel A corresponds to Station BA functional mean, the red color curve (dotted line) to Station CO, and the blue one (dashed line) to station UV. Panel B shows the pointwise p-values (solid black line); the dashed light blue line represents the level of significance of the test.

For the yearly global estimation, a functional analysis of variance test was performed to gain a general idea of air pollution in the three stations. Results are shown in Fig. 2 Fig. 2A depicts the $PM_{2.5}$ functional means at each of the three stations; Fig. 2B shows the pointwise p-value for the functional analysis of variance, from which we concluded that $PM_{2.5}$ concentrations show significant differences for hours from 0 to 15 on an average day at all three assessed stations.

Analysis with smoothing parameters per day and per station

Environmental pollution could be influenced by the day of the week due to the impact of mobile sources because the circulation of internal combustion vehicles is consistently higher during some days of the week. To evaluate the results in light of this fact, a day-to-day analysis was performed. Fig. 3 reveals that air pollution on an average Monday resembles that of the entire year (Fig. 2). This behavior is similar on Tuesdays, Saturdays and Sundays; but differs on Wednesdays, Thursday and Fridays, as shown in Fig. 4D for Thursdays, where the curves between stations are not significant during most of the day. In fact, even though significant differences were identified during the first hours of the day, atypical observations occur during those hours,

corresponding to the first hours of the year, especially during New Year celebrations, as shown in Fig. 4 A and B.

These results lead to the conclusion that the three stations behave differently, as observed in the global analysis for the year 2015, confirming the importance of taking air quality readings in all three sites.

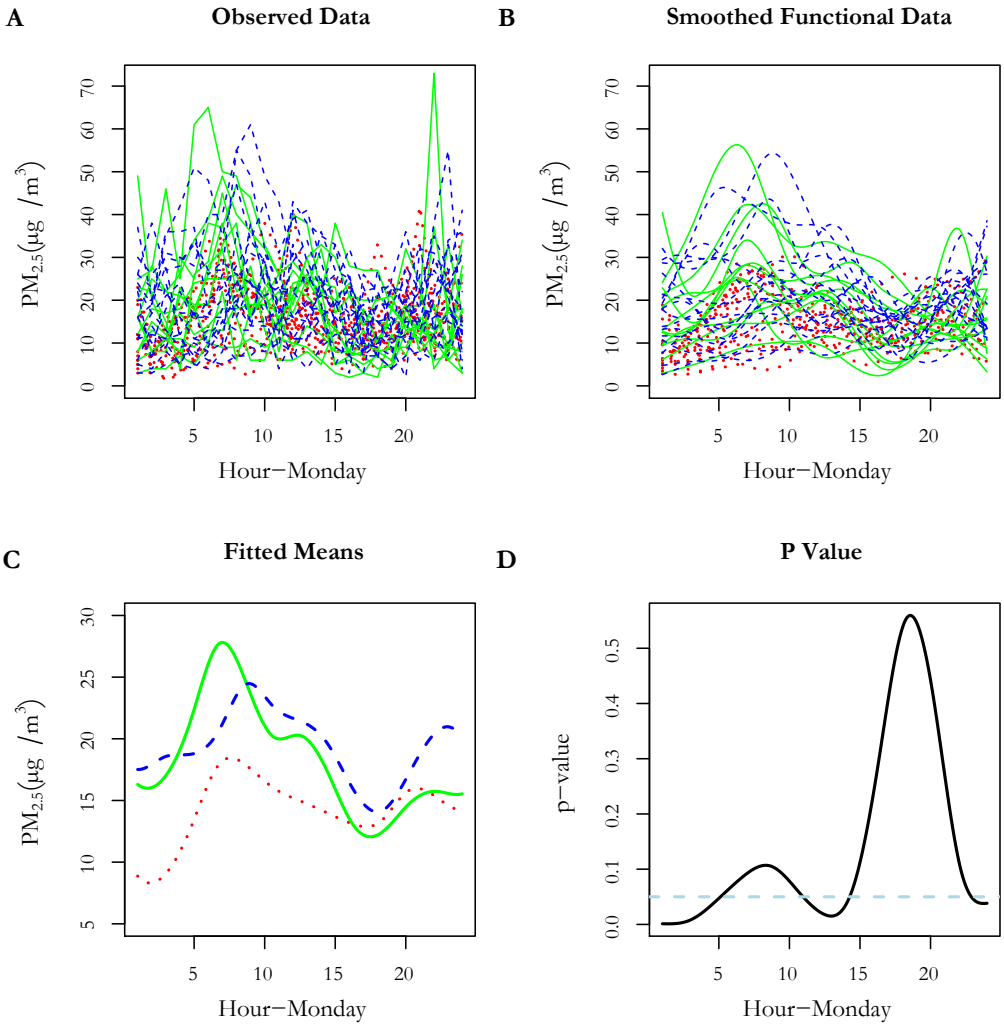

Figure 3. Results for the construction of the curves and the functional analysis of variance for Mondays. Observed and Functional data are shown in panels A and B, respectively. Green color is used for Station BA, red for Station CO, and blue for Station UV. Fig. 3C displays the stations functional means and panel D the pointwise p-values (solid black line); the dashed light blue line represents the level of significance of the test.

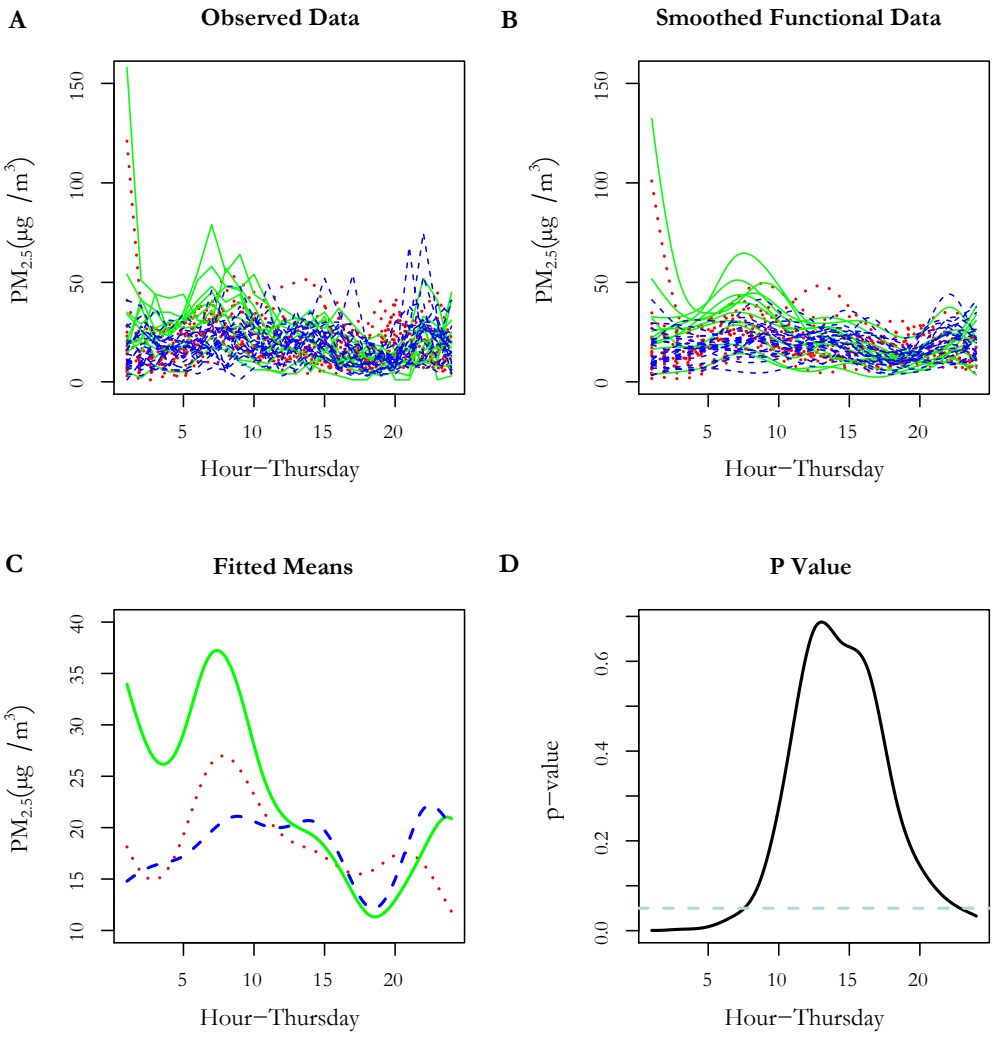

Figure 4. Results for the construction of the curves and the functional analysis of variance for Thursdays. Observed data and Functional data are shown in panels A and B. Green color is used for Station BA, red for Station CO, and blue for Station UV. Panel C displays the stations functional means and panel D the pointwise p-values (solid black line); the dashed ligth blue line represents the level of significance of the test.

Conclusions

- The environmental pollution data, when measured in an infinite and continuous space, in this case time, turn out to be appropriate for the analysis of functional data. This analysis to summarize in a smooth curve the information of n scalar values of a variable that varies continuously and is discretely observed.

- The results obtained through the data exploration are reflected in the functional analysis of variance because in general, environmental pollution in the three monitoring stations shows significant differences noticeable during morning hours and almost imperceptible during afternoon-evening hours.
- Thanks to this analysis, we determined that the station with the highest average pollution levels per day is station BA. Consequently, the inhabitants of this zone, which is to the north-east of the city, are exposed to a higher risk of contracting chronic respiratory diseases. Revealling the need for implementing environmental policies that to reduce $PM_{2.5}$ levels in such zones of the city.
- According to the present functional exploratory analysis (Fig. 1), stations CO and UV seem to show similar average $PM_{2.5}$ levels, even if local environmental conditions differ from one station to another. A further analysis is needed on this issue.
- Due to the large amount of missing data, the functional data were reduced. An option to tackle this drawback would be the imputation of missing data to complete the dataset. In this way, a greater number of curves will be available for analysis, since only days with complete hourly measurements are considered in the present work.
- Finally, as a matter of priority, we recommended to evaluate the correlations between measurements of different stations, which could be high because the studied pollutant can be airborne. This suggests a possible spatial and temporal correlation between monitoring stations. In addition, in order for the correlation to have greater meaning, measurements should be taken on the same dates at the studied stations, although this implies a reduction in the database.

Acknowledgements

The authors would like to especially thank the Colombian Administrative Department of Science, Technology and Innovation (Colciencias) and Universidad del Valle, which made this work possible through the 2015 Young Researchers Program and research grant C.I. 2842 “Statistical modelling of air pollution due to particles of aerodynamic diameter less than $2.5\mu m$ ($PM_{2.5}$)”.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

Al-Hamdan M, Crosson WL, Limaye AS, Rickman DL, Quattrochi DA, Estes Jr MG, Qualters JR, Sinclair AH, Tolsma D, Adeniyi KA, Niskar AS. Methods for characterizing fine particulate matter using ground observations and remotely sensed data: Potential use for environmental public health surveillance. *Journal of the Air Waste Management Association*, 2009.

doi: 10.3155/1047-3289.59.7.865

Bell M, Dominici F, Ebisu K, Zeger S, Samet J. Spatial and temporal variation in PM2.5 chemical composition in the united states for health effects studies. *Environmental Health Perspectives*, 7(115): 989-995, July 2007.

doi: 10.1289/ehp.9621

Estévez-Pérez G, Vilar J. Functional anova starting from discrete data: an application to air quality data. *Environmental and Ecological Statistics*, 20(3): 495-517, September 2013. ISSN 1573-3009.

doi: 10.1007/s10651-012-0231-2

Febrero-Bande M, Galeano P, González-Manteiga W. A functional analysis of NOx levels: location and scale estimation and outlier detection. *Computational Statistics*, 2007.

doi: 10.1007/s00180-007-0048-x

Febrero-Bande M, De la Fuente MO. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4), October 2012.

doi: 10.18637/jss.v051.i04

Ferraty F, Vieu P. *Nonparametric Functional Data Analysis Theory and Practice*. Springer Series in Statistics. Springer, 2006.

doi: 10.1007/0-387-36620-2

Górecki T, Smaga Ł. Multivariate analysis of variance for functional data. *Journal of Applied Statistics*, 44(12): 2172-2189, 2017.

doi: 10.1080/02664763.2016.1247791

Górecki T, Smaga Ł. fdANOVA: An R software package for analysis of variance for univariate and multivariate functional data. *Computational Statistics*, 2018.

doi: 10.1007/s00180-018-0842-7

Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, 2nd ed, 2009.

doi: 10.1007/BF02985802

James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

doi: 10.1007/978-1-4614-7138-7

Laden F, Neas LM, Dockery DW, Schwartz J. Association of fine particulate matter from different sources with daily mortality in six u.s. cities. *Environmental Health Perspectives*, 108(10), October 2000.

doi: 10.1289/ehp.00108941

Ramsay JO, Silverman BW. *Functional Data Analysis*. Springer, New York, 2nd ed, 2005.

doi: 10.1007/b98888

Ramsay JO, Wickham H, Graves S, Hooker S. *Package "fda"*. R Project, July 2018.

https://cran.r-project.org/web/packages/fda/fda.pdf

Ruiz-Medina MD. Functional analysis of variance for hilbert-valued multivariate fixed effect models. *Statistics*, 50(3): 689-715, 2016.

doi: 10.1080/02331888.2015.1094069

Zhang JT. *Analysis of Variance for Functional Data*. Chapman Hall, London, 2013.

Análisis de varianza funcional de la contaminación del aire causada por partículas finas

Resumen: La contaminación ambiental es perjudicial para la salud humana, ya que puede conducir a enfermedades respiratorias crónicas. En particular, las partículas finas suspendidas en el aire ($PM_{2.5}$) se cuentan entre los contaminantes atmosféricos más agresivos. Los niveles de $PM_{2.5}$ varían según las condiciones locales. El objetivo de este trabajo fue comparar las lecturas de $PM_{2.5}$ aerotransportadas realizadas durante todo el año en tres estaciones de vigilancia de la calidad del aire en Cali (Colombia) para determinar si estas muestran una variación espacial y temporal significativa. Sometimos el conjunto de datos $PM_{2.5}$ obtenido a un análisis de varianza funcional. Observamos que los niveles de $PM_{2.5}$ varían significativamente entre los tres sitios de medición en una escala temporal. Mientras que en las horas de la mañana los niveles de $PM_{2.5}$ en estos tres sitios diferían más, en las horas de la tarde y la noche los niveles de $PM_{2.5}$ correspondientes no fueron significativamente diferentes.

Palabras clave: partículas en el aire; contaminación ambiental; datos funcionales.

Análise de variância funcional da poluição do ar causada por partículas finas

Resumo: A poluição ambiental é prejudicial à saúde humana, pois pode levar a doenças respiratórias crônicas. Em particular, partículas finas suspensas no ar ($PM_{2.5}$) contam entre os poluentes atmosféricos mais agressivos. Os níveis de $PM_{2.5}$ variam dependendo das condições locais. O objetivo deste trabalho foi comparar as leituras de $PM_{2.5}$ no ar realizadas ao longo do ano em três estações de vigilância da qualidade do ar em Cali (Colômbia) para determinar se elas mostram variação espacial e temporal significativa. Submetemos o conjunto de dados $PM_{2.5}$ obtido a uma análise de variância funcional. Observamos que os níveis de $PM_{2.5}$ variam significativamente entre os três locais de medição em escala temporal. Enquanto nas horas da manhã os níveis de $PM_{2.5}$ nesses três locais diferiam mais, nas horas da tarde e da noite, os níveis correspondentes de $PM_{2.5}$ não eram significativamente diferentes.

Palavras-chave: partículas em suspensao; poluição ambiental; dados funcionais.

Javier Olaya Ochoa

Statistician from the Universidad del Valle (Cali, Colombia) in 1985, MSc in Mathematical Sciences (1997), and PhD in Management Science (2000) from Clemson University (Clemson, SC, USA). Currently works as Full Professor of Statistics in the School of Statistics at the Universidad del Valle, Cali, Colombia.

ORCID: 0000-0001-7014-2782

Diana Paola Ovalle Muñoz

Statistician from the Universidad del Valle (Cali, Colombia) in 2014. Currently is a Master in Statistics' student at the same University.

ORCID: 0000-0002-5408-5762

Cristhian Leonardo Urbano León

Mathematician from the Universidad del Cauca (Popayán, Colombia) in 2012. MSc in Statistics (2019) from Universidad del Valle (Cali, Colombia).

ORCID: 0000-0003-4622-538X