DESCOMPONIENDO PARTÍCULAS SUSPENDIDAS

ENCONTRANDO PATRONES TEMPORALES Y ESPACIALES DE CONCENTRACIÓN DE PM2.5 EN LA CDMX MEDIANTE MFPCA

Lars Daniel Johansson Niño
Instituto Tecnológico Autónomo de México

Resumen

Las partículas PM2.5 son un tipo de partículas suspendidas —productos de la quema de combustibles para generar energía, emisiones vehiculares e industriales, etc.— cuya inhalación está asociada a enfermedades respiratorias, cáncer de pulmón, entre otros efectos adversos. Con el fin de proteger la salud pública el Sistema de Monitoreo Atmosférico (SIMAT) cuenta con estaciones distribuidas en la zona metropolitana de la CDMX que registran las concentraciones de PM2.5 de forma horaria. En este trabajo se utilizaron estos registros para construir curvas semanales de concentración de PM2.5 para 19 estaciones del SIMAT. Posteriormente se ajustó un modelo funcional mixto estimado por medio de un Análisis Funcional de Componentes Principales Multinivel (MFPCA). Los componentes del modelo permitieron identificar patrones geográficos y temporales entre las curvas de las estaciones.

Datos

Los datos corresponden a registros horarios de 19 estaciones del SIMAT, que se muestran en el mapa de la Fig. 1, descargados de [4]. Para la estación $i \in \{1, 2, ..., 19\}$, llame $y_{ij,k}$ al registro de la hora $k = 1, 2, ..., 168$ y la semana $j = 1, 2, ..., 51$. Desde una perspectiva de Análisis de Datos Funcionales (FDA) cada $y_{ij,k}$ puede verse un punto observado de una curva continua latente $Y_{ij}(t)$ de concentración semanal de PM2.5. i.e. $Y_{ij}(t_k) = y_{ij,k}$ para $t_k \in \{1, 2, ..., 168\}$. Así, la colección $\{y_{ij,k}\}$ da evidencia de una muestra de funciones $Y_{ij}(t)$ cuyo comportamiento se espera esté asociado a la posición geográfica de la estación i y la semana j en la que fue registrada.

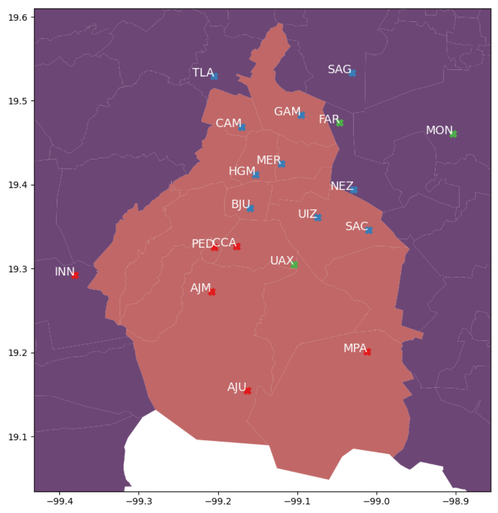

Fig. 1: 19 estaciones del SIMAT. Los colores corresponden a grupos identificados por medio de la Fig. 2

Hemos de notar que, para algunas ij's, existen horas k's que no fueron observadas. Si bien el MFPCA implementado permite la existencia de datos faltantes, solo se utilizaron las ij's que tenían al menos un 60% de sus puntos observados. Por esto mismo, no se realizó ningún proceso de interpolación o ajuste a los datos.

Análisis Funcional de Componentes Principales Multinivel (MFPCA)

Debido a que múltiples curvas provienen de una estación y periodo temporal, es necesario modelar la dependencia inducida para obtener resultados luables. El MFPCA hace esto expresando a cada $Y_{ij}(t)$ como una desviacion aleatoria de una media global $\mu(t)$ determinada por 4 componentes: $\eta_j(t)$, un común a todas las i's aplicables; $U_i(t)$, un efecto aleatorio de estación; $V_{ij}(t)$ un efecto de la semana j especifica a la estación i que explica lo que los otros componentes no explican, y $\varepsilon_{ij}(t)$, un error de observación. Esto nos deja con el modelo funcional mixto (FMM) (1).

$$Y_{ij}(t) = \mu(t) + \eta_j(t) + U_i(t) + V_{ij}(t) + \varepsilon_{ij}(t) \quad (1)$$

Lo que diferencia al MFPCA de otros modelos mixtos es que $U_i(t)$ y $V_{ij}(t)$ se expresan como una suma aleatoria de funciones ortnormales, un análogo funcional al PCA tradicional.

$$c_U(s,t) = \sum_{k\geq 1} \lambda_k \phi_k(s)\phi_k(t) \quad c_V(s,t) = \sum_{k\geq 1} \nu_k \varphi_k(s)\varphi_k(t) \quad (2)$$

En particular, existen bases ortonormales $\{\phi_k(t)\}$ e $\{\varphi_k(t)\}$ y valores propios $\lambda_1 \geq \lambda_2 \geq ...$ e $\nu_1 \geq \nu_2 \geq ...$ de forma que la función covarianza de U_i, $c_U(s,t) = cov\{U_i(s), U_i(t)\}$ y de V_{ij}, $c_V(s,t) = cov\{V_{ij}(s), V_{ij}(t)\}$ satisfacen (2). Además, las $U_i(t)$'s y $V_{ij}(t)$'s se pueden expresar como una combinación lineal de las $\{\phi_k(t)\}$ e $\{\varphi_k(t)\}$ cuyos coeficientes son v.a's. $\{\xi_{ik}\}$ e $\{\zeta_{ijk}\}$. Estas, llamadas scores por tradición, son de media cero, no correlacionadas, y tales que $Var(\xi_{ik}) = \lambda_k$ e $Var(\zeta_{kij}) = \nu_k$ dándonos (3).

$$U_i(t) = \sum_{k\geq 1} \xi_{ik}\phi_k(t) \quad V_{ij}(t) = \sum_{k\geq 1} \zeta_{ijk}\varphi_k(t) \quad (3)$$

Al igual que en PCA, uno puede hablar de la Proporción de Variación explicada (PVE). En este caso, se habla por la PVE por el nivel 1, correspondiendo a las $U_i(t)$'s y el nivel dos, correspondiente a las $V_{ij}(t)$'s. e.g. el primero está dado por $(\sum_{k\geq 1}\lambda_k)/(\sum_k \lambda_k + \sum_k \nu_k)$. También se puede hablar de una PVE especifica al nivel. En particular, $\lambda_{r,\phi}/\sum_k \lambda_k$ y $\nu_{r'}/\sum_k \nu_k$ dan la PVE por $\phi_r(t)$ e $\varphi_{r'}(t)$ en el primer y segundo nivel respectivamente. Los componentes del modelo fueron estimados por medio de la función mfpca.face del paquete refund [3] de R.

Análisis espacial

Las $U_i(t) = \sum_{k\geq 1} \xi_{u,ik}\phi_k(t)$ y sus componentes, ya que representan los efectos de estación, nos darán indicio de patrones espaciales en las estaciones. Al igual que en el PCA, las posiciones relativas de los scores $\xi_{u,ik}$ ayudan a detectar que observaciones, en este caso estaciones, tienen comportamientos similares. Los scores de las funciones $\phi_1(t)$ e $\phi_2(t)$, que explican alrededor del 91% de la variabilidad del nivel 1, revelan patrones geograficos al analizarles por cuadrantes de $\mathbb{R}^2$.

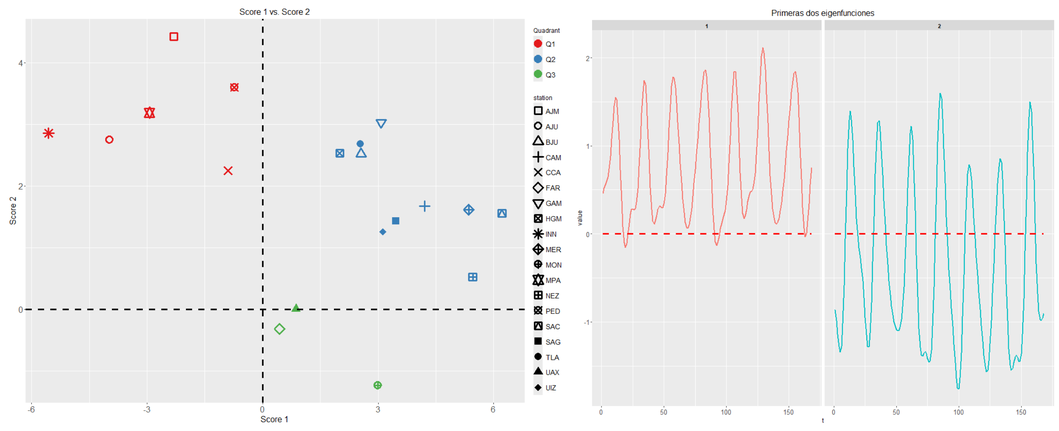

Fig. 2: **Izquierda:** $\xi_{u,i1}$ vs $\xi_{u,i2}$ estimadas. **Derecha:** $\phi_1(t)$ y $\phi_2(t)$ estimadas. Estas explican alrededor del 91% de la variabilidad del primer nivel.

De particular importancia es el signo de los scores. Para un $t_0 \in [0, 168]$, si $\phi_k(t_0) > 0$ (< 0) y $\xi_{ki} > 0$ entonces significa que $\phi_k(t_0)$ contribuye a $Y_{ij}(t_0)$ de forma positiva (negativa). Esto es, si $\xi_{ik} > 0$, $\phi_k(t_0)$ contribuye a $Y_{ij}(t_0)0)$ o acorde a su signo. Bajo un razonamiento analogo, si $\xi_{ki} < 0$ entonces $\phi_k(t)$ contribuye a $Y_{ij}(t_0)$ de forma contraria a su signo. En la parte izquierda de Fig. 2 se identifican tres grupos en base a ξ_{1i} e ξ_{2i}: el **rojo**, que incluye las estaciones con $\xi_{1i} < 0$ y $\xi_{2i} > 0$; el **azul**, con scores que cumplen $\xi_{1i}, \xi_{2i} > 0$, y el **verde** cuyos scores son tales que $\xi_{1i} > 0$ y $\xi_{2i} < 0$. Como se muestra en la Fig. 1 los grupos mencionados marcan un patrón regional claro en la CDMX. El grupo azul esta compuesto primordialmente por estaciones del coloquialmente llamado *norte de la CDMX*, mientras el rojo por el *sur* de la ciudad en su mayoría.

Los efectos de estación $U_i(t)$ completos, impresos en la Fig. 3 y colorados acorde a los cuadrantes de la Fig. 2, ayudan a identificar si la i-ésima estación tiene una contaminación mas arriba o abajo del promedio. En concreto, el que $U_i(t_0) > 0$ (< 0) en t_0 significa que la i-ésima estación, sin considerar otros componentes, tiene una concentración mas arriba (abajo) que el promedio en el punto t_0 de la semana.

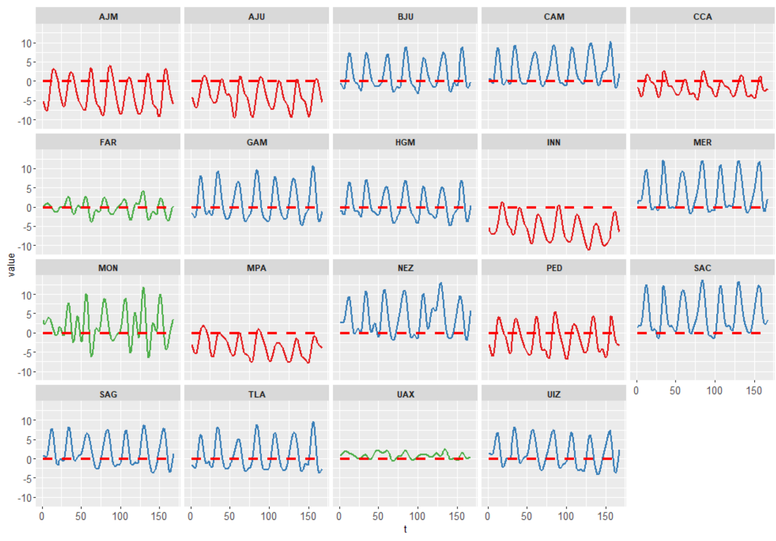

Fig. 3: $U_i(t)$'s estimadas por estación coloreadas acorde a sus grupos de la Fig. 2

Como se observa en la Fig. 3, las estaciones pertenecientes al grupo rojo tienden a tener una contaminación más **abajo** del promedio, mientras las del azul una más **arriba**. Por tanto, los patrones detectados por cuadrantes también *dividen* a las estaciones que, por lo general, tienen una contaminación más arriba/abajo al promedio. Esto es de esperarse, pues sus ξ_{1i} e ξ_{2i} son *similares* por grupo ademas de que $\phi_1(t)$ y $\phi_2(t)$ explican buena parte de la variación dentro del nivel 1.

Una vez mencionado lo ultimo, vale la pena observar que el primer nivel solo explica alrededor del 10.6% de la variación total de las $Y_{ij}(t)$'s. e.g. $\sum_{k\geq 1}\lambda_{k,\phi}/(\sum_k \lambda_{k,\phi} + \sum_k \lambda_{k,\varphi}) \approx 0.106$. Lo cual permite concluir que, si bien esta distinción de grupos si existe, no es cercana a ser suficiente para explicar el comportamiento de las curvas en su totalidad.

Análisis temporal

Para el análisis temporal utilizaremos las $\eta_j(t)$ ya que estos representan efectos fijos semanales. Al igual que antes, el que $\eta_j(t_0) > 0$ $(<)$ significa que, en el punto t_0, la semana j tiene una contribución positiva (negativa) sobre las $Y_{ij}(t_0)$'s. En la Fig. 4 el efecto estimado $\eta_j(t)$ aparece en el cuadro de un mes particular si la semana j comienza en tal mes.

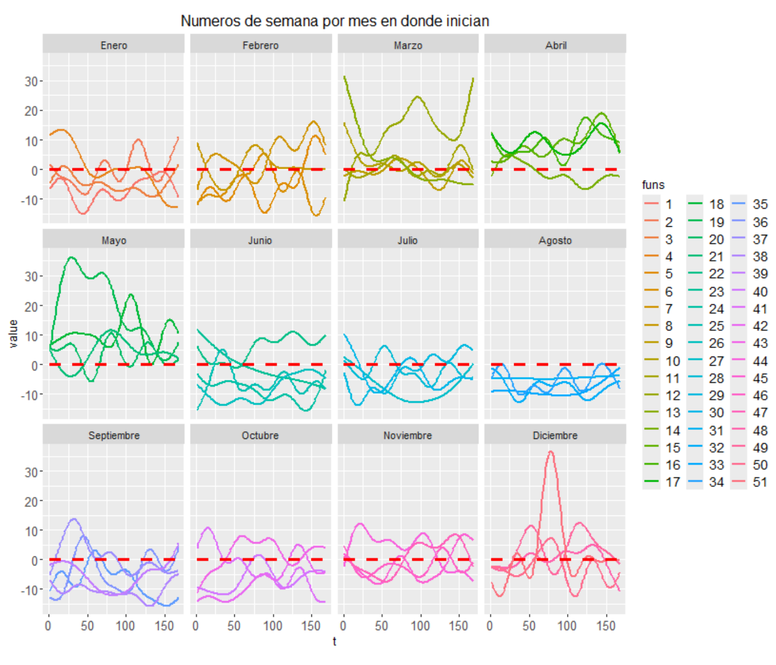

Fig. 4: $\eta_j(t)$ estimadas.

La Fig. 4 exhibe las $\eta_j(t)$'s de algunos meses tienen un tanto efectos generales positivos y negativos sobre las $Y_{ij}(t)$'s (e.g. Noviembre, o Marzo). Por otro lado, las $\eta_j(t)$'s de meses como Enero, Junio, Julio, Agosto, Septiembre, y Octubre, a ojo, tienen efectos generalmente mas abajo del promedio. También podría argumentarse que las funciones de Abril, Mayo, y quizás Marzo tienen un nivel de contaminación mayor al promedio.

Vale la pena mencionar que, aunque no está incluido en este análisis, el segundo nivel, el de las $V_{ij}(t)$'s, tiene un rol importante al modelar las $Y_{ij}(t)$'s. El PVE de este nivel, que incluye información especifica a la semana por cada estación, fue de casi 0.9. Sus scores, sin embargo, no mostraron evidencia de patrones evidentes específicos a las estaciones o semanas.

Conclusiones

En este análisis se exhibió que los componentes estimados del modelo (1) con (3) permiten apreciar patrones regionales y temporales. Las estaciones del grupo azul, primordialmente del norte, tendieron a tener una concentración mayor a las del grupo rojo, con mayoría en el sur, y el verde. En adición, las semanas que iniciaron en enero, junio, julio, agosto, septiembre y octubre tuvieron una concentración menor al promedio mientras que las que inician en Abril, Mayo, y Marzo una mayor. Aun con esto, mas allá del aspecto descriptivo, el que la mayoría de la variación estuviera explicada por el segundo nivel muestra que modelar la concentración de PM2.5 exige considerar factores locales y temporales específicos.

Referencias

[1] Comisión Ambiental de la Megalópolis. *Partículas Suspendidas: Características y Principales Fuentes*. Gobierno de México. 2018. URL: https://www.gob.mx/comisionambiental/articulos/particulas-suspendidascaracteristicas-y-principales-fuentes?idiom=es (visited on 05/03/2025).

[2] Ciprian M. Crainiceanu et al. *Functional Data Analysis with R*. CRC Press, 2024. DOI: 10.1201/9781003278726.

[3] Jeff Goldsmith et al. *refund: Regression with Functional Data*. R package version 0.1-36, commit 1562aeee0e6ec4188306bbcd626bcbfb1d119cd8. 2024. URL: https://github.com/refunders/refund.

[4] Secretaría del Medio Ambiente de la Ciudad de México. *Concentraciones de contaminantes atmosféricos – Sistema de Monitoreo Atmosférico (SIMAT)*. 2025. URL: http://www.aire.cdmx.gob.mx/estadisticas-consultas/concentraciones/index.php (visited on 05/03/2025).