Decomposing Suspended Particles

Finding temporal and spatial patterns on PM2.5 concentration levels in Mexico City using MFPCA

Lars Daniel Johansson Niño
Instituto Tecnológico Autónomo de México

Abstract

PM2.5 particles are a type of suspended particulate matter—byproducts of burning fuel for energy, vehicle and industrial emissions, etc.—whose inhalation is associated with respiratory illnesses, lung cancer, and other adverse health effects.To protect public health, the Atmospheric Monitoring System (SIMAT, by its Spanish acronym) has stations distributed throughout the metropolitan area of Mexico City (CDMX) that record hourly PM2.5 concentrations. This work used these records to construct weekly PM2.5 concentration curves for 19 SIMAT stations. Subsequently, a functional mixed model was fitted, estimated through Multilevel Functional Principal Component Analysis (MFPCA). The model's components allowed for the identification of geographic and temporal patterns among the stations' curves.

Data

The data correspond to hourly records from 19 SIMAT stations, shown on the map in Fig. 1, downloaded from [4]. For station $i \in 1,2,...,19$, let $y_{ij,k}$ be the record for hour $k = 1,2,...,168$ and week $j = 1,2,...,51$. From a Functional Data Analysis (FDA) perspective, each $y_{ij,k}$ can be viewed as an observed point of a latent continuous curve $Y_{ij}(t)$ of weekly PM2.5 concentration, i.e., $Y_{ij}(t_k) = y_{ij,k} + \varepsilon_{ij,k}$ for $t_k \in 1,2,...,168$. Thus, the collection $y_{ij,k}$ provides evidence of a sample of functions $Y_{ij}(t)$ whose behavior is expected to be associated with the geographical location of station i and the week j in which it was recorded.

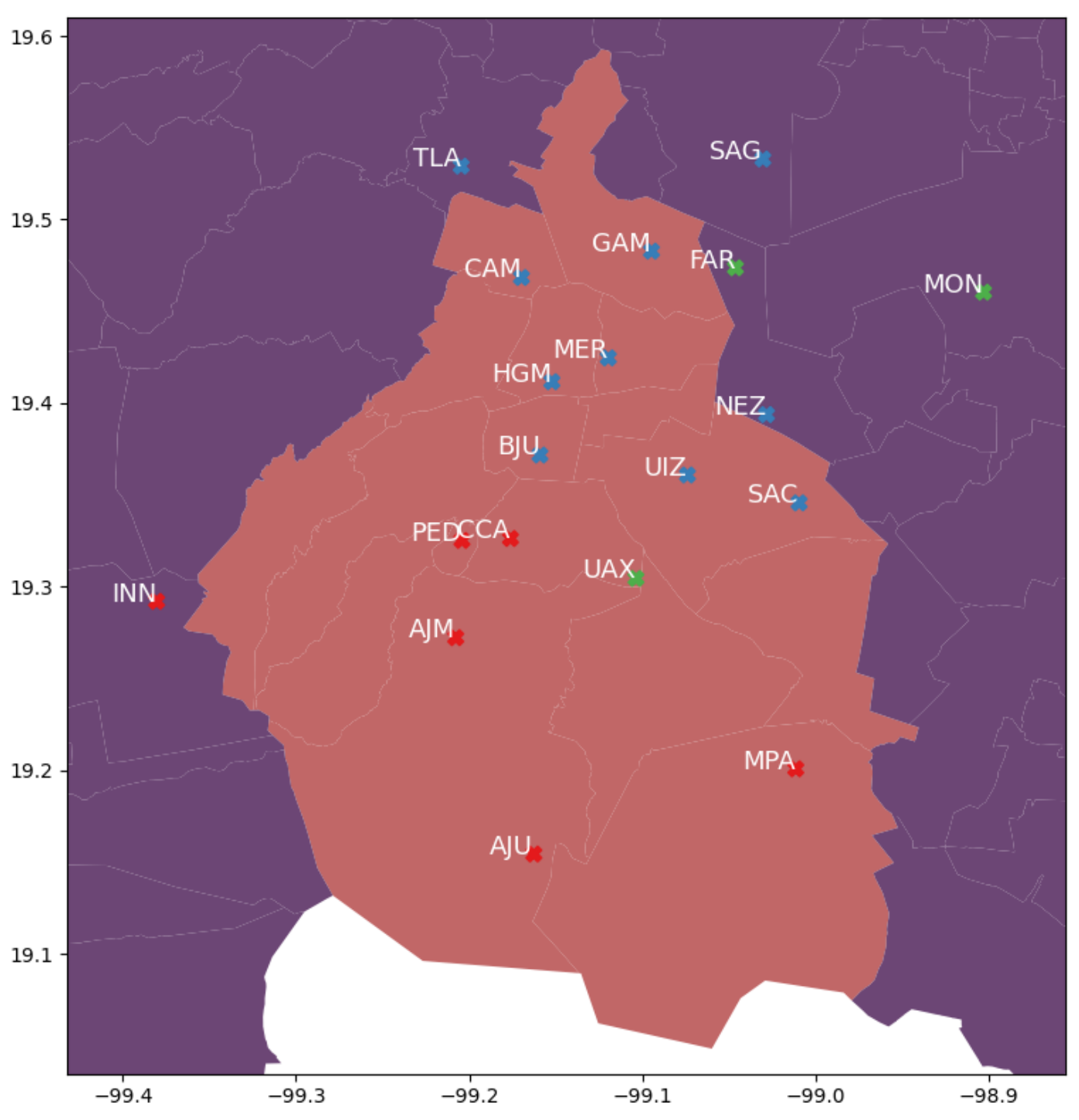

Fig. 1: 19 stations from SIMAT. The colours correspond to the identified groups from Fig. 2

It should be noted that for some ij's, there are hours k's that were not observed. Although the implemented MFPCA allows for the existence of missing data, only those ij's that had at least 60% of their points observed were used. For this same reason, no interpolation or data fitting process was carried out.

Multilevel Functional Principal Component Analysis (MFPCA)

Because multiple curves come from a single station and time period, it is necessary to model the induced dependence to obtain reliable results. The MFPCA accomplishes this by expressing each $Y_{ij}(t)$ as a random deviation from a global mean $\mu(t)$ determined by 4 components: $\eta_j(t)$, a week specific fixed effect common to all applicable i's; $U_i(t)$, a station-specific random effect; $V_{ij}(t)$ an effect of week j specific to station i that accounts for within subject variation, and $\varepsilon_{ij}(t)$, an observation error. This leaves us with the functional mixed model (FMM) (1).

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

What differentiates MFPCA from other mixed models is that $U_i(t)$ and $V_{ij}(t)$ are expressed as a random sum of orthonormal functions, a functional analogue to traditional PCA.

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

In particular, there exist orthonormal bases $\phi_k(t)$ and $\varphi_k(t)$ and eigenvalues $\lambda_1 \geq \lambda_2 \geq ...$ and $\nu_1 \geq \nu_2 \geq ...$ such that the covariance function of U_i, $c_U(s,t) = cov\{U_i(s), U_i(t)\}$ and of V_{ij}, $c_V(s,t) = cov\{V_{ij}(s), V_{ij}(t)\}$ satisfy (2). Furthermore, the $U_i(t)$'s and $V_{ij}(t)$'s can be expressed as a linear combination of the $\phi_k(t)$ and $\varphi_k(t)$ whose coefficients are random variables ξ_{ik} and ζ_{ijk}. These, traditionally called scores, have zero mean, are uncorrelated, and are such that $Var(\xi_{ik}) = \lambda_k$ and $Var(\zeta_{ijk}) = \nu_k$ giving us (3).

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

As in PCA, one can refer to the Proportion of Variance Explained (PVE). In this case, one refers to the PVE for level 1, corresponding to the $U_i(t)$'s, and level two, corresponding to the $V_{ij}(t)$'s. For example, the first is given by $(\sum_{k\geq 1} \lambda_k)/(\sum_k \lambda_k + \sum_k \nu_k)$. One can also refer to a PVE specific to a level. In particular, $\lambda_r/\sum_k \lambda_k$ and $\nu_{r'}/\sum_k \nu_k$ give the PVE by $\phi_r(t)$ and $\varphi_{r'}(t)$ in the first and second level, respectively. The model components were estimated using the `mfpca.face` function from the `refund` [3] package in R.

Spatial analysis

The $U_i(t) = \sum_{k\geq 1} \xi_{u,ik}\phi_k(t)$ and their components, since they represent the station effects, will give us an indication of spatial patterns in the stations. As in PCA, the relative positions of the scores $\xi_{u,ik}$ help to detect which observations, in this case stations, have similar behaviors. The scores of the functions $\phi_1(t)$ and $\phi_2(t)$, which explain around 91% of the level 1 variability, reveal geographic patterns when analyzed by quadrants of $\mathbb{R}^2$.

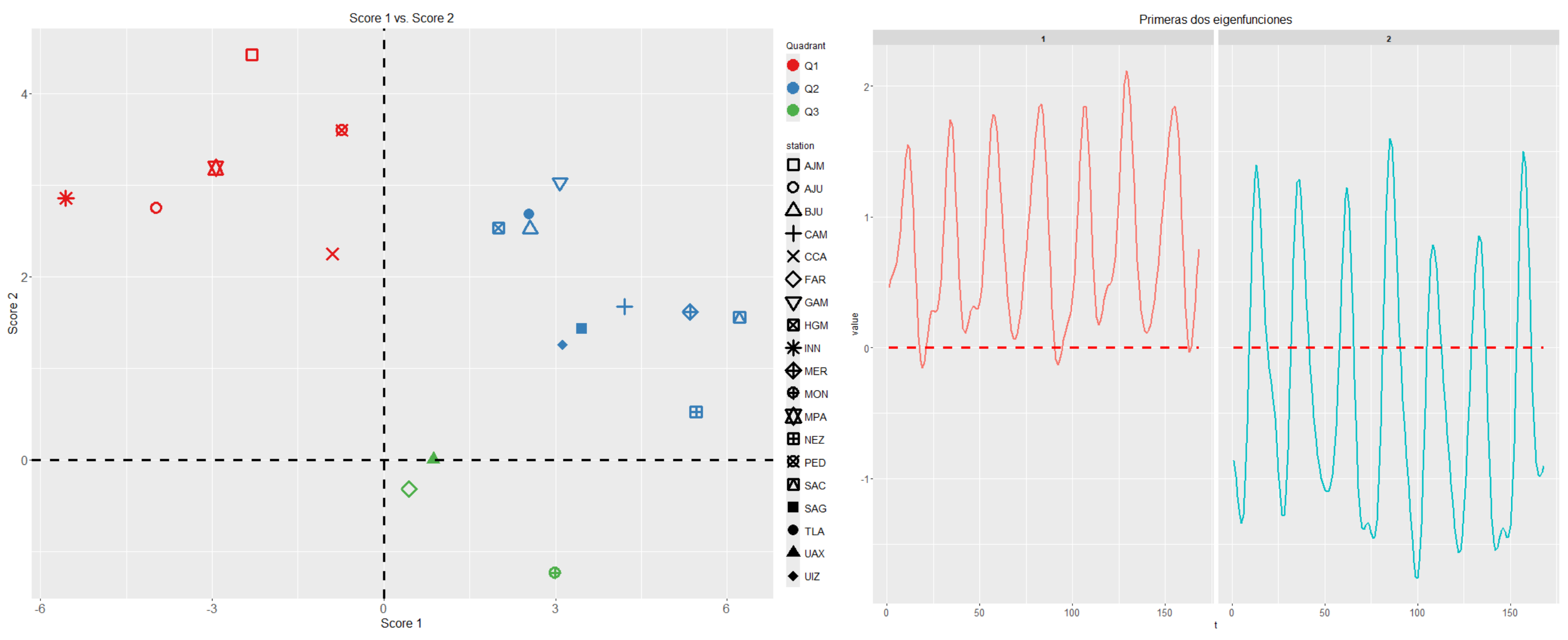

Fig. 2: **Left:** $\xi_{u,i1}$ vs $\xi_{u,i2}$ estimations. **Right:** $\phi_1(t)$ y $\phi_2(t)$ estimations. The previous explain around 91% of level 1 variability.

Of particular importance is the sign of the scores. For a $t_0 \in [0,168]$, if $\phi_k(t_0) > 0$ (< 0) and $\xi_{ki} > 0$ then it means that $\phi_k(t_0)$ contributes to $Y_{ij}(t_0)$ in a positive (negative) way. That is, if $\xi_{ik} > 0$, $\phi_k(t_0)$ contributes to $Y_{ij}(t_0)$ in accordance with its sign. Under analogous reasoning, if $\xi_{ki} < 0$ then $\phi_k(t)$ contributes to $Y_{ij}(t_0)$ in a way contrary to its sign. On the left side of Fig. 2, three groups are identified based on ξ_{1i} and ξ_{2i}: the **red** one, which includes stations with $\xi_{1i} < 0$ and $\xi_{2i} > 0$; the **blue** one, with scores satisfying $\xi_{1i}, \xi_{2i} > 0$, and the **green** one whose scores are such that $\xi_{1i} > 0$ and $\xi_{2i} < 0$. As shown in Fig. 1, the aforementioned groups mark a clear regional pattern in Mexico City (CDMX). The blue group is primarily composed of stations from the colloquially called *north of Mexico City*, while the red one mostly consists of stations from the *south* of the city.

The full station effects $U_i(t)$, shown in Fig. 3 and colored according to the quadrants of Fig. 2, help to identify whether the i-th station has pollution levels above or below the average. Specifically, $U_i(t_0) > 0$ (< 0) at t_0 means that the i-th station, without considering other components, has a concentration above (below) the average at the point t_0 of the week.

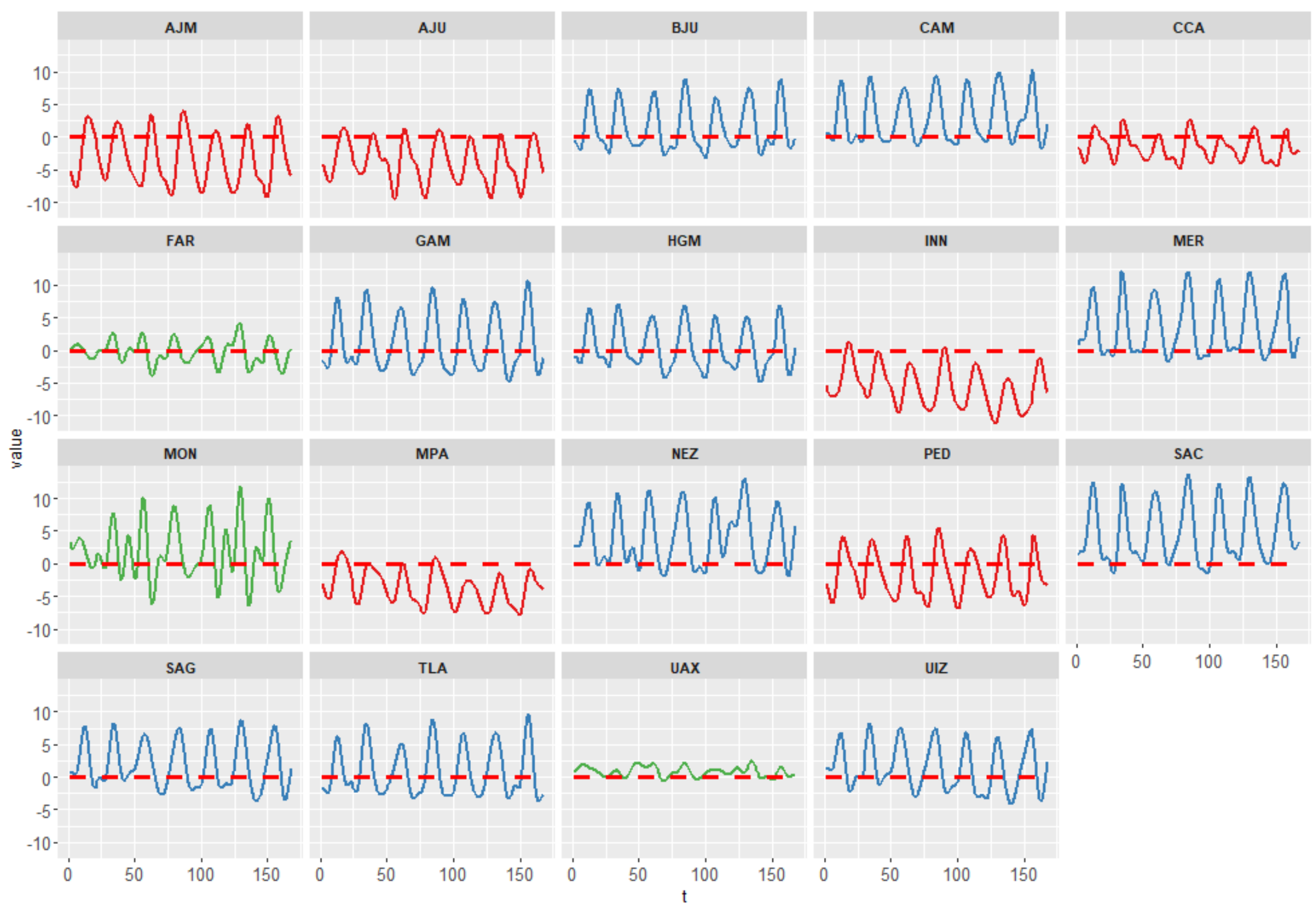

Fig. 3: $U_i(t)$'s estimations. The colors correspond to the identified groups via Fig. 2

As can be seen in Fig. 3, the stations belonging to the red group tend to have pollution levels **below** the average, while those in the blue group have levels **above** it. Therefore, the patterns detected by quadrants also *divides* the stations that generally have pollution levels above/below the average. This is to be expected, as their ξ_{1i} and ξ_{2i} are *similar* within each group, and because $\phi_1(t)$ and $\phi_2(t)$ explain a large portion of the variation within level 1.

Having mentioned the latter, it is worth noting that the first level only explains about 10.6% of the total variation of the $Y_{ij}(t)$'s, e.g., $\sum_{k\geq 1} \lambda_{k,\phi}/(\sum_k \lambda_{k,\phi} + \sum_k \lambda_{k,\varphi}) \approx 0.106$. This allows us to conclude that, although this group distinction does exist, it is far from sufficient to explain the behavior of the curves in their entirety.

Temporal Analysis

For the temporal analysis, we will use the $\eta_j(t)$ since these represent fixed weekly effects. As before, if $\eta_j(t_0) > 0$ (< 0) it means that, at the point t_0, week j has a positive (negative) contribution to the $Y_{ij}(t_0)$'s. In Fig. 4, the estimated effect $\eta_j(t)$ appears in the box of a particular month if week j begins in that month.

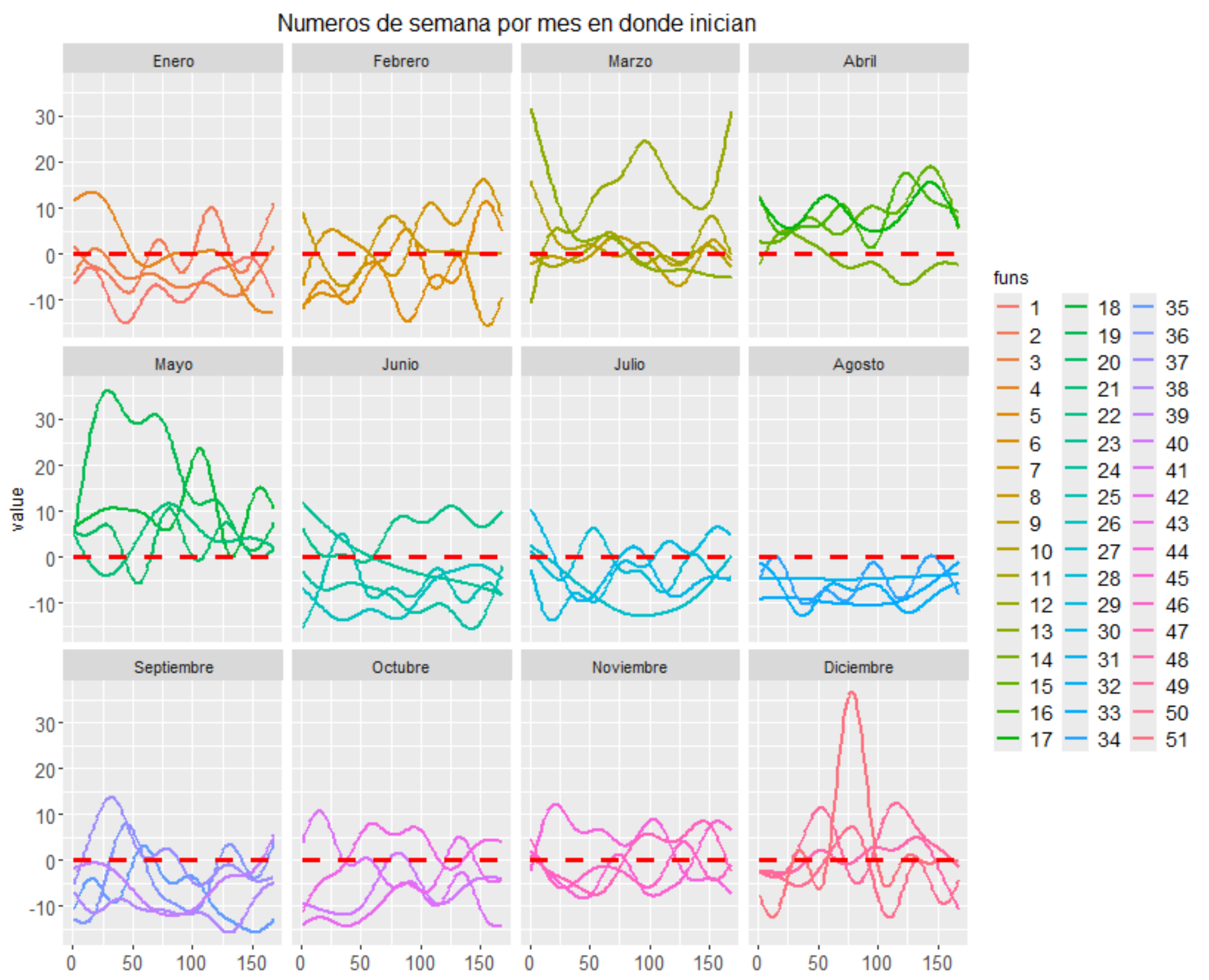

Fig. 4: $\eta_j(t)$ estimations.

Fig. 4 shows that the $\eta_j(t)$'s of some months have both positive and negative general effects on the $Y_{ij}(t)$'s (e.g., November or March). On the other hand, the $\eta_j(t)$'s of months such as January, June, July, August, September, and October, appear to have effects generally below the average. It could also be argued that the functions for April, May, and perhaps March have a pollution level higher than the average.

It is worth mentioning that, although not included in this analysis, the second level, that of the $V_{ij}(t)$'s, plays an important role in modeling the $Y_{ij}(t)$'s. The PVE of this level, which includes information specific to the week for each station, was almost 0.9. Its scores, however, did not show evidence of clear patterns specific to the stations or weeks.

Conclusions

This analysis revealed that the estimated components of model (1) with (3) allow for the observation of regional and temporal patterns. The stations in the blue group, primarily from the north, tended to have a higher concentration than those in the red group—which were mostly in the south—and the green group. In addition, the weeks starting in January, June, July, August, September, and October had a concentration lower than the average, while those starting in April, May, and March had a higher one. Even so, beyond this descriptive aspect, the fact that most of the variation was explained by the second level shows that modeling PM2.5 concentration requires the consideration of specific local and temporal factors.

References

[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).