Functional data analysis of PM 2.5 level in Beijing from 2010-2014

Yixing Lu[1]

[1] Biostatistics Graduate Group, UC Davis

Introduction

- **Beijing** had serious chronic **air pollution** during 2010-2014, especially PM 2.5 problems.
- **PM 2.5** are inhalable particulate matters with diameter smaller than 2.5 um.
- Epidemiology evidence has shown that prolonged exposure to high level of PM 2.5 could **cause lung morbidity, respiratory diseases and lung cancer**. [1]
- **Meteorological parameters** such wind speed, wind direction, precipitation, temperature, etc. could highly influence daily PM 2.5 levels.
- There was a **seasonal change** in PM 2.5 levels in Beijing.

Objectives

- Use **non-parametric smoothing** and **functional principal component analysis (FPCA)** to unveil the seasonal patterns of PM 2.5 level and wind speed in Beijing.
- Use FPCA and functional linear regression (FLR) to model the relationship between **wind speed** and **daily average PM 2.5 level.**

Methods

Pre-processing

- Omitted days with completely **missing PM 2.5 data**
- Took **daily average** of PM 2.5 and wind speed values
- Divide days into four **seasons,** and regard each season of each year as a functional observation

Smoothing

- **Fourier basis smoothing with roughness penalty**
- $y_{ij} = f_i(t_{ij}) + \varepsilon_{ij} = \sum_{k=1}^{K} c_k B_k(t_{ij}) + \varepsilon_{ij}$
- Normal equation: $\widetilde{Q}(c_1, \dots, c_k) = \sum_{i=1}^{N} \left\{ \sum_{j=1}^{T_i} [y_{ij} - f_i(t_{ij})]^2 + \kappa \int_{\Gamma} [f_i^{"}(t)]^2 \, dt \right\}$

FPCA

- **Input smoothed curves as noiseless data** into FPCA
- $f_i^L(t_{ij}) = \mu(t_{ij}) + \sum_{l=1}^{L} \xi_{il} \phi_l$

FLR

- **Regress first two FPCA scores of PM 2.5 onto those of wind speed**
- $E(Y(t)|X) - \mu_Y(t) = \int \beta(s,t)[X(s) - \mu_X(s)]ds$
- $\beta(s,t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \beta_{mn} \phi_m(t) \psi_n(s)$

Results

Serious PM 2.5 pollution in Beijing 2010-2014

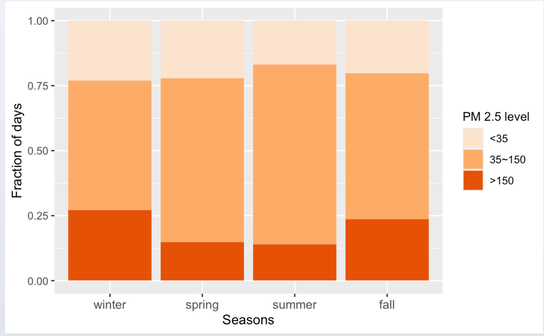

Figure 1: Fraction of days under three PM 2.5 level categories during different seasons: 2010-2014 Beijing.

Smoothed PM 2.5 curves by seasons

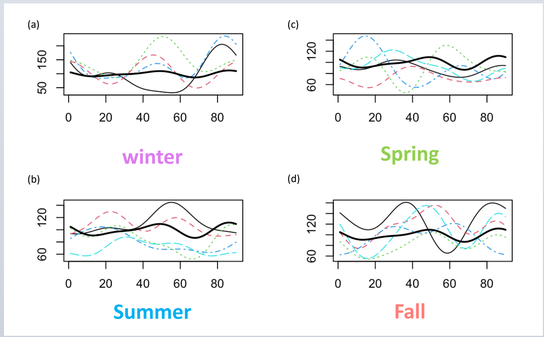

Figure 2: Smoothed PM 2.5 curves grouped by seasons with common mean curve in bold black line. (a) winter, (b) summer, (c) spring, (d) fall.

Smoothed Wind Speed curves by seasons

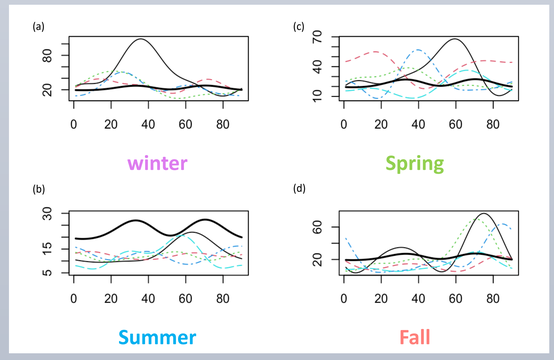

Figure 3: Smoothed wind speed curves grouped by seasons with common mean curve in bold black line. (a) winter, (b) summer, (c) spring, (d) fall.

FPCA results for PM 2.5

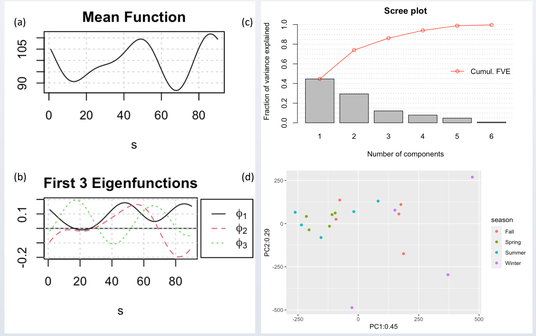

Figure 4: FPCA results summary for PM 2.5. (a) Mean function; (b) First three eigenfuncions; (c) Scree plot; (d) Score plot for the first 3 PCs.

FPCA results for Wind Speed

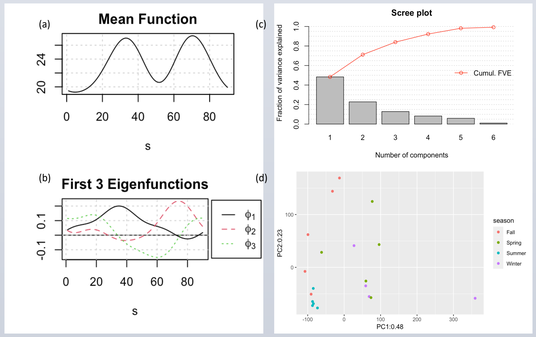

Figure 5: FPCA results summary for wind speed. (a) Mean function; (b) First three eigenfuncions; (c) Scree plot; (d) Score plot for the first 3 PCs.

FLR coefficient function surface plot

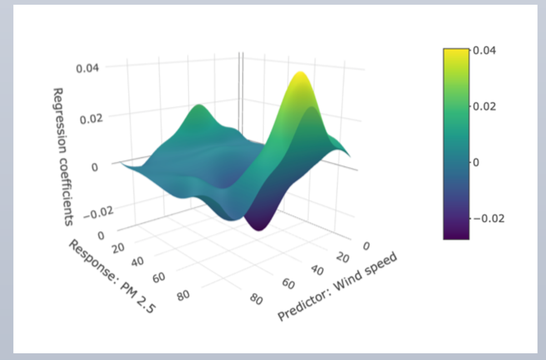

Figure 6: Surface plot of functional linear regression coefficient for PM 2.5 vs. wind speed.

Discussion

Smoothing:

- Both PM 2.5 and wind speed curves **had large seasonal variability**.
- PM 2.5 curves had larger variance in the middle and towards the end of each season, while wind speed had larger variance during early-mid season.
- **Mean functions** of PM 2.5 and wind speed had **opposite trends**, implying a potential negative association.

Decomposition of estimated covariance into eigenfunctions:

- **PM 2.5:**
  - PC1 (45% variance) separated spring and winter curves
  - PC2 (29% variance) deepened the valley around day 70
- **Wind speed:**
  - PC1 (48% variance) explained the peak at day 30, PC2 (23% variance) explained the peak at day 70-80.
  - PC1 separated winter from summer and fall, where winter curves peaked during earlier period of a season while summer and fall peaked at the later period.

Functional linear regression:

- Most of the time wind speed was negatively associated with PM 2.5
- The most substantial effect of wind speed on reducing PM 2.5 appeared between mid-season wind speed and late-season PM 2.5.

Conclusion

- Seasonal pattern differences were identified for both PM 2.5 level and wind speed using smoothing and FPCA.
- Mean function of PM 2.5 and wind speed showed opposite trend and FLR confirmed they were negatively associated most of the time.
- Without including other predictors and confounder, no causal interpretation could be made.

References

- Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., & Ito, K. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution JAMA 287 (9): 1132-1141.
- US-EPA. (2012). Revised air quality standards for particle pollution and updates to the Air Quality Index (AQI).
- Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257.
- Xiao, W., & Hu, Y. (2018, July). Functional data analysis of air pollution in six major cities. In Journal of Physics: Conference Series (Vol. 1053, No. 1, p. 012131). IOP Publishing.
- Muller, HG. STA 223 lecture notes. Winter 2021

Acknowledgement

- Dataset used for analysis in the current project was retrieved from UCI machine learning repository: http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data
- The dataset and working R code could be found at: https://github.com/yixlu/FunctionalDataAnalysis_PM2.5.

Functional data analysis of PM 2.5 level in Beijing from 2010-2014

Yixing Lu, Biostatistics Graduate Group, UC Davis

Abstract

Beijing experienced serious air pollution, especially PM 2.5 problems during 2010-2014. The daily PM 2.5 levels could be affected by season, weather, and some other meteorological parameters. This project used functional data analysis and time series data on PM 2.5 in Beijing to unveil the seasonal change in PM 2.5 levels and its relationship with wind speed, which is believed to alleviate air pollution to some extent. The data smoothing and functional principal component analysis revealed differences in PM 2.5 patterns among different seasons and an opposite trend to that of wind speed. Further functional linear regression showed that wind speed was negatively associated with PM 2.5 level most of the time.

1. Introduction

Beijing and many major cities in China experienced serious chronic air pollution in the past ten years. The problem with particulate matter, especially PM 2.5, had raised concern among general population because of its environmental and health hazardous impact. Epidemiology evidence showed that prolonged exposure to PM 2.5 at high levels could cause lung morbidity, serious respiratory diseases, and even cancer [1]. According to National Ambient Air Quality Standards (NAAQs) for PM, the 24-hour standard for PM 2.5 is 35 ug/m^3 [2], above which is considered unhealthy to sensitive people. In Beijing during 2010-2014, only around 20% of days had low or moderate level of PM 2.5 [3]. PM 2.5 level is highly influenced by meteorological parameters, seasons, and many other factors. In this project, functional data analysis was used to investigate the time-series data of PM 2.5 level and wind speed in Beijing during 2010-2014. *The main objectives are to find the seasonal differences in patterns of PM 2.5 level and its relationship with seasonal wind speed pattern.*

2. Data source

The dataset analyzed in this project was provided to UCI Machine Learning Repository by Song Xi Chen (http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data). Hourly PM 2.5 data was recorded at UC Embassy in Beijing (116.47E, 39.95N), while hourly meteorological measurements were taken at Beijing Capital International Airport, obtained from weather.nocrew.org [3]. The two locations were 17km apart and was considered to have the same air quality and weather conditions. The time-series data run from January 1st, 2010 to December 31st, 2014, a total of 5 years' data. Besides PM 2.5, wind speed was also used in this project to investigate its relationship with PM 2.5. Days with completely missing PM 2.5 readings were omitted from the data set, and the daily average values of both PM 2.5 and wind speed was used for analysis. Daily average values were subject to functional data analysis to unveil the change in seasonal patterns of Beijing PM 2.5 level and its relationship with wind speed. After removing missing values and taking daily average, there were 1788 values for both variables in chronical order.

3. Methodology

3.1. Fourier basis smoothing with roughness penalty

In the context of functional data analysis, both the PM 2.5 concentration and wind speed data are assumed to be continuous. Since both variables are subject to measurement errors, they can be expressed as $y_{ij} = f_i(t_{ij}) + \varepsilon_{ij}$, where $i = 1, \dots, N, j = 1, \dots, T_i, t_{ij} \in \Gamma$. The variable is the sum of a continuous function defined within a time interval domain Γ and a random disturbance factor ε_{ij}. $f_i(t_{ij})$ can be expanded as a linear combination of basis functions: $f_i(t_{ij}) = \sum_{k=1}^{K} c_k B_k(t_{ij})$, where ϕ_k is the basis function and c_k is the coefficient. Fourier basis is often used for periodic data, which takes the following form: $B_0 = 1, B_{2r-1} = \sin(rwt), B_{2r} = \cos(rwt)$, where $r = \frac{K-1}{2}$, K is the number of basis. In order to accommodate for the measurement error, a penalty term was included. The coefficients $c_1, \dots, c_k$ are found by minimizing the sum of squared fitting residuals with roughness penalty: $\tilde{Q}(c_1, \dots, c_k) = \sum_{i=1}^{N}\{\sum_{j=1}^{T_i}[y_{ij} - f_i(t_{ij})]^2 + \kappa \int_{\Gamma}[f_i''(t)]^2 dt\}$, where κ is the penalty parameter which was chosen based on minimizing the sum of generalized cross validation (GCV) of all curves and subjective inspection of the smoothing to achieve a balance between bias and variance. The smoothed data were used as the input for downstream function principal component analysis.

3.2. Functional principal component analysis (FPCA)

After smoothing the data, our observations became 19 random curves: each corresponded to a season of a year. Since winter was defined as December to February and 2009 December data was not available, the curve corresponding to winter 2010 was excluded. Random functions as observations gave rise to the necessity of dimension reduction, which is commonly done by functional principal component analysis. FPCA extracted information from the random curves by generating functional principal components (FPC) for each curve. The smoothed curves were regarded as noise-less, and could be expressed as: $f_i(t_{ij}) = \mu(t_{ij}) + \sum_{l=1}^{\infty} A_{il}\phi_l$, where ϕ_l are orthogonal eigenfunctions and A_{il} are corresponding scores. In reality, it can be truncated to the first L components and estimated as: $f_i^L(t_{ij}) = \mu(t_{ij}) + \sum_{l=1}^{L} \xi_{il}\phi_l$. Since the input data into FPCA were regarded noiseless and dense, the mean and covariance of the curves were estimated by cross-sectional method as specified in the "FPCA" function in "fdapace" R package.

3.3. Functional liner regression

After conducting FPCA on both PM 2.5 and wind speed smoothed data, the first m and n principal components and corresponding eigenfunctions of the response and predictor variables were used for functional linear regression. The normal equation of the regression model is defined as: $E(Y(t)|X) - \mu_Y(t) = \int \beta(s,t)[X(s) - \mu_X(s)]ds$. With the following representation of the response and predictor as in the FPCA model: $Y(t) = \mu_Y(t) + \sum_{m=1}^{\infty} A_m\phi_m(t), X(s) = \mu_X(s) + \sum_{n=1}^{\infty} B_n\psi_n(s)$,

where A_m and B_n are the functional principal components (scores) while ϕ_m and ψ_n are eigenfunctions for the response and predictor respectively, we get $\beta(s,t) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\frac{E[B_nA_m]}{E[B_n^2]}\phi_m(t)\psi_n(s) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\beta_{mn}\phi_m(t)\psi_n(s)$, where β_{mn} is the coefficient when regressing the m^{th} PC score of the response variable onto the n^{th} PC score of the predictor.

4. Results

4.1. Smoothing of seasonal PM 2.5 data and wind speed data

The number of basis used in Fourier basis smoothing was chosen to be 7 which could capture the important trends in the curves. The penalty parameter was evaluated in respect to the sum of GCV for all curves. ***Figure S1*** showed the sum of GCV as a function of log (κ) for smoothing of both seasonal PM 2.5 and wind speed. The penalty parameters that minimized the sum of GCV for each variable were chosen to smooth the curves. The smoothed curves with original discrete observation points could be found in ***Figure S2-S3***. Mean curves and 95% point-wise confidence bands for PM 2.5 and wind speed were shown in ***Figure 2***, while the curves grouped by seasons were shown in **Figure 3**.

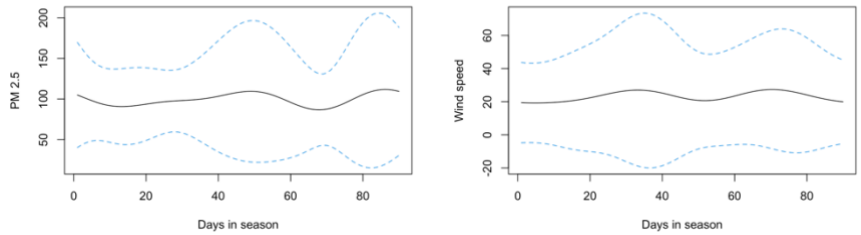

Figure 2: Mean smoothed curve and 95% point-wise confidence band. Left: PM 2.5, right: wind speed.

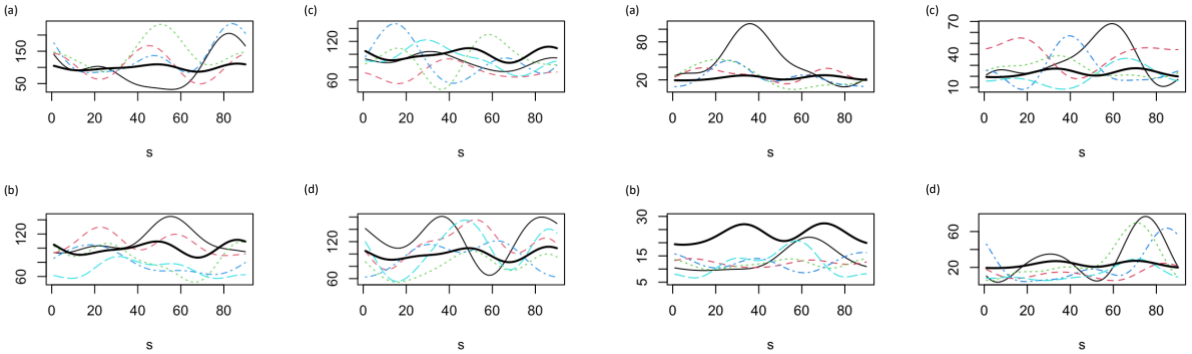

Figure 3: Smoothed curves grouped by seasons with common mean curve in bold black line. (a) winter, (b) summer, (c) spring, (d) fall. Left 4 panels: PM 2.5, right 4 panels: wind speed.

4.2. FPCA on seasonal PM 2.5 and wind speed

The smoothed curves were subject to FPCA, results shown in ***Figure 4*** for PM 2.5, and ***Figure 5*** for wind speed. The first two principal components of PM 2.5 accounted for 74% of total variance, and the first two principal components of wind speed accounted for 71% of total variance (***Figure 4c, 5c***). Also, by comparing the mean function (***Figure 4a, 5a***) and the decomposition of the estimated covariance

surface into the first two components (***Figure 4b, 5b***), we could see that the first two PCs accounted for the major peaks and valleys in the mean function curve. The score plots (***Figure 4d, 5d***) in the space of the first two PCs were also shown here, while those in PC1-PC3 and PC2-PC3 space could be found in Appendix ***Figure S4*** and ***S5***.

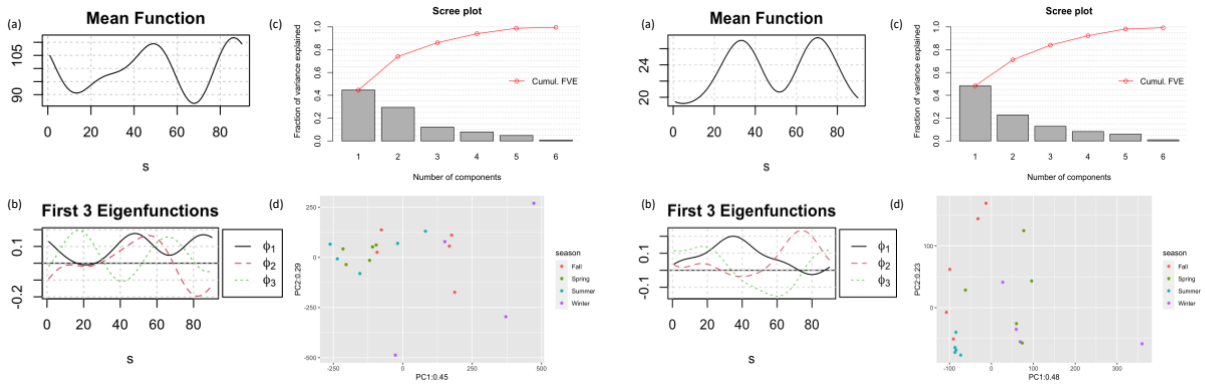

Figure 4: FPCA results summary for PM 2.5 (left panel), and wind speed (right panel). *(a) Mean function; (b) First three eigenfuncions; (c) Scree plot; (d) Score plot for the first 3 PCs.*

4.3. FLR of seasonal PM 2.5 on wind speed

Comparing the mean function of PM 2.5 and that of wind speed shown in ***Figure 4 and 5,*** we could see that in general, the local minimum of PM 2.5 occurred around the similar time as the local maximum in wind speed, indicating that there might be a negative association between PM 2.5 concentration and wind speed. In order to test the null hypothesis that there is no significant linear regression effect, a functional liner regression was conducted using the scores of the first two principal components of both PM 2.5 (response) and wind speed (predictor). Pairwise linear regression between the four scores were performed and the estimated coefficients (β_{mn}) were shown in ***Table S1***. To test the assumption of linear relationship between the response and predictor variables, residuals of each liner regression model were plotted against fitted values and the plots could be found in Appendix in ***Figure S5***. $\beta(s,t)$ was calculated and the surface plot was shown in ***Figure 6***.

5. Discussion

The shapes of the smoothed mean curves of PM 2.5 and wind speed in ***Figure 2*** indicated that both variables had substantial within-season variation. Mean PM 2.5 concentration appeared to be higher at the beginning, middle, and end of the seasons, while there was a local minimum during middle-late period. Also, there appeared a shoulder in the curve around the time of one month into the seasons. On the predictor side, wind speed had the opposite pattern to PM 2.5 that the mean wind speed was higher one month after the season began

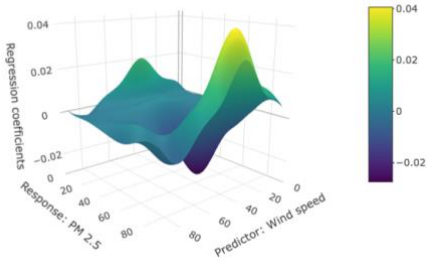

Figure 6: Surface plot of functional linear regression coefficients.

and half month before the season ended. From the 95% point-wise confidence band, we could see that both variables had large variance, due to the difference among four seasons. PM 2.5 curves had larger variance in the middle and towards the end of each season, while wind speed had larger variance during early-mid season.

Figure 4 and 5 showed the estimated covariance surface decomposition into the first three eigenfunctions for PM 2.5 and wind speed respectively, as well as the score plots. For PM 2.5, the first eigenfunction in general had the same trend as the mean curve, while the second eigenfunction deepened the valley between 60 and 80 days, and the third eigenfunction mainly contributed to the shoulder between day 20 and day 40. For wind speed, the first two eigenfunctions accounted for the peak at day 30, and day 70 respectively, while the third eigenfunction for the valley in between. The score plots provided information on how close each functional observation was correlated with the principal components. In ***Figure 4d,*** the score plot showed that three of the four winter curves had the largest positive scores on PC 1, while spring curves were on the other negative half of the axis, indicating these two seasons had largest difference in their pattern of PM 2.5 concentration in the direction specified by PC 1. Comparing the winter and spring curves in ***Figure 3***, we could see that the peak of PM 2.5 concentration appeared in latter half of winter but early half of spring. Considering the two seasons are in sequence, it was expected that there was a continuous pattern extended from winter to spring. In terms of wind speed, PC 1 mainly separated winter from spring and summer, where winter had higher wind speed during the earlier half of the season, in line with what the first eigenfunction showed, while spring and summer peaked at latter half of the season. Since the PM 2.5 concentration and the wind speed seemed to have opposite trends, a functional linear regression was conducted to test whether such association existed. Among the four pairwise simple linear regressions, only the one regressed PC 2 of PM 2.5 on PC 1 of wind speed had a significant negative coefficient (***Table S1***). The coefficient function surface plot in ***Figure 6*** showed that most of the time wind speed was negatively associated with PM 2.5, except that wind speed at around one month into the seasons appeared to have positive association with PM 2.5 level at very beginning and very end of the seasons. The most substantial effect of wind speed on reducing PM 2.5 appeared between mid-season wind speed and late-season PM 2.5.

6. Conclusion

Non-parametric data smoothing together with FPCA revealed seasonal changes in patterns of PM 2.5 level in Beijing and the opposite trend in wind speed. Result of functional linear regression confirmed that wind speed was negatively associated with PM 2.5 most of the time. However, given the presence of other confounder, such as precipitation, wind direction, no causal inference could be made between wind speed and PM 2.5 level.

Reference

1. Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., & Ito, K. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution JAMA 287 (9): 1132–1141.
2. US-EPA. (2012). Revised air quality standards for particle pollution and updates to the Air Quality Index (AQI).
3. Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257.
4. Xiao, W., & Hu, Y. (2018, July). Functional data analysis of air pollution in six major cities. In Journal of Physics: Conference Series (Vol. 1053, No. 1, p. 012131). IOP Publishing.

Appendix

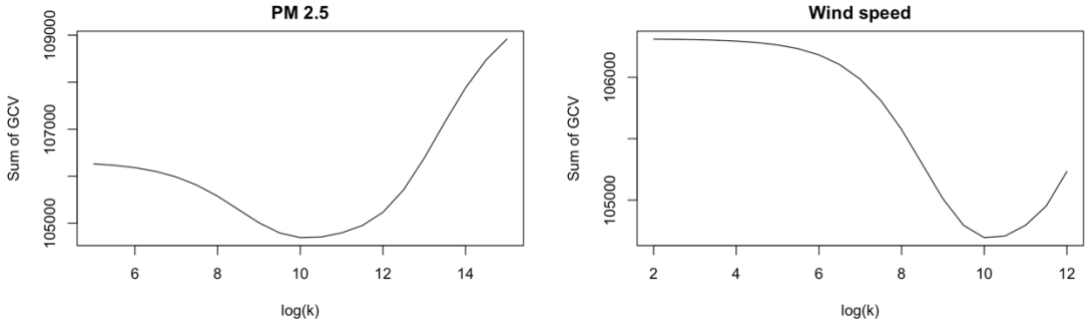

Figure S1: Sum of GCV vs. log (κ). *Penalty parameter minimized the sum of GCV was chosen to smooth the curves.*

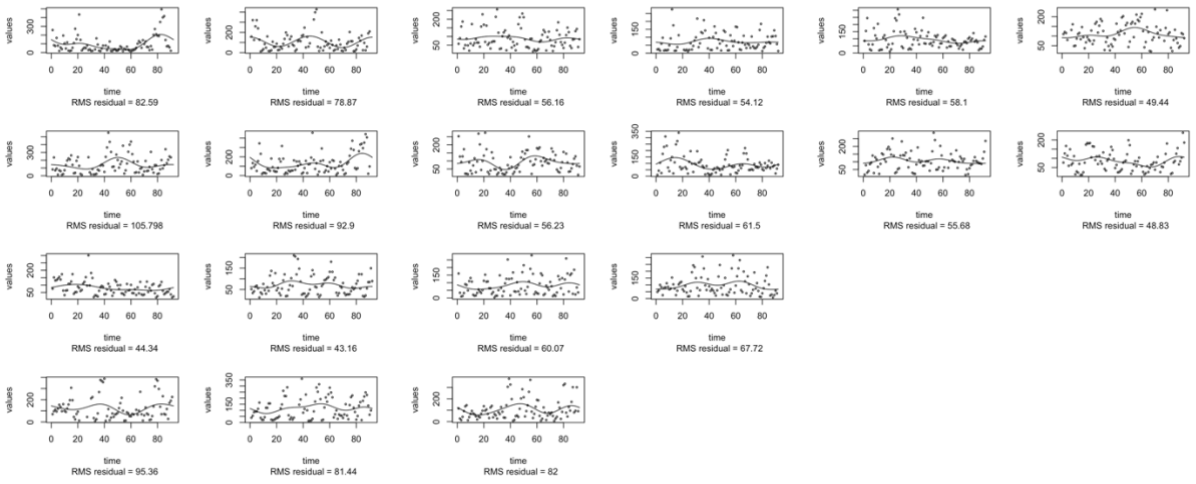

Figure S2: Smoothed curves of seasonal PM 2.5 using Fourier basis.

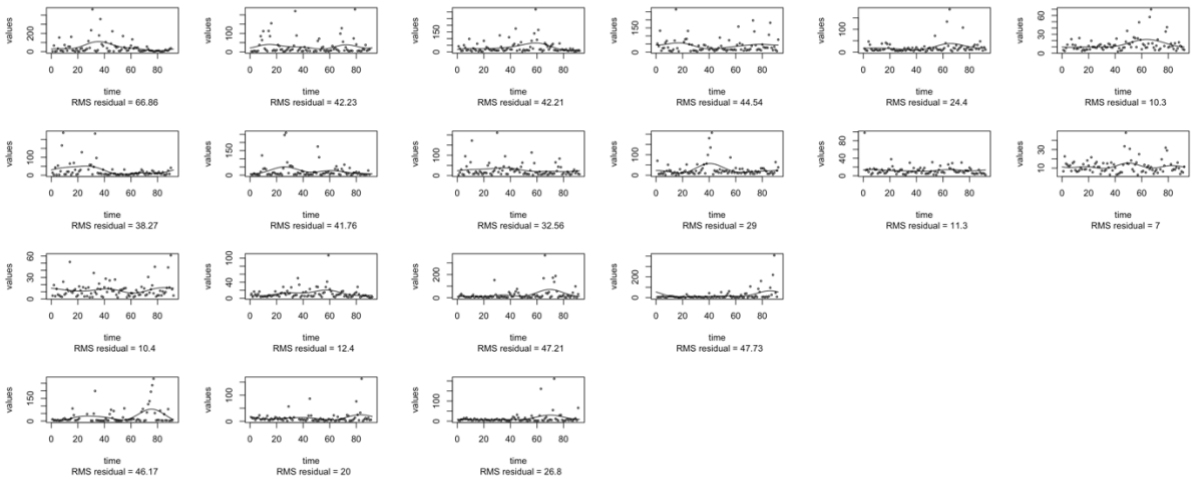

Figure S3: Smoothed curves of seasonal wind speed using Fourier basis.

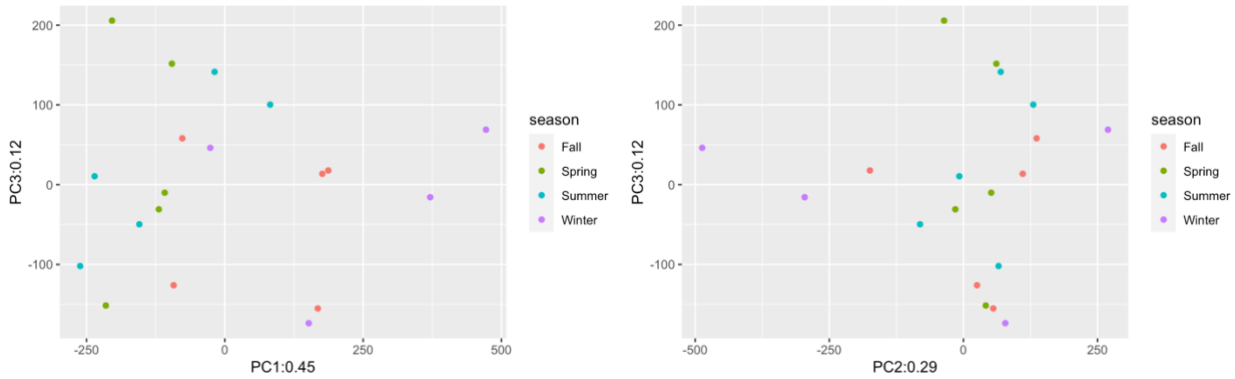

Figure S4: Score plots in PC1-PC3 (left) and PC2-PC3 (right) space for PM 2.5.

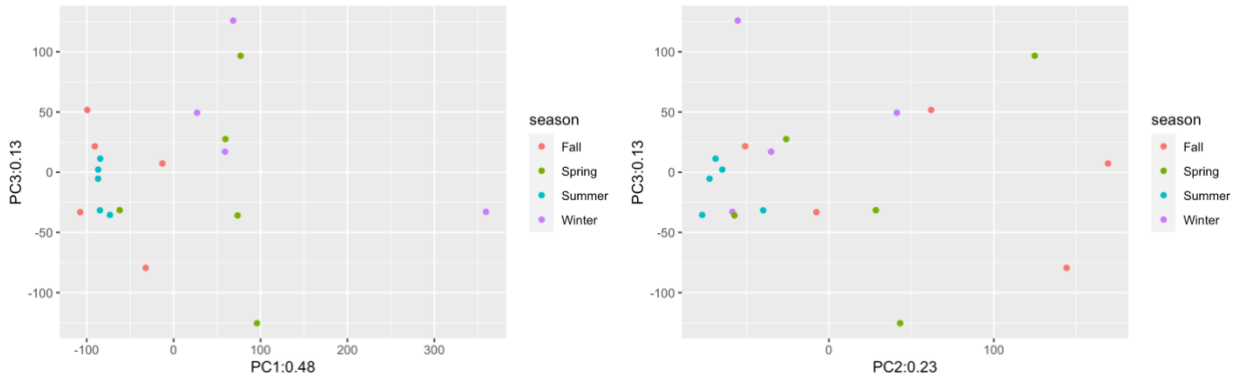

Figure S5: Score plots in PC1-PC3 (left) and PC2-PC3 (right) space for wind speed.

Table S1: Estimated coefficients in linear regression of PM 2.5 scores on wind speed scores. Response: PM 2.5

		PC 1		PC 2	
		β	p-value	β	p-value
Predictor:	**PC 1**	0.12	0.79	-0.93	0.004*
wind speed	**PC 2**	-0.03	0.96	0.018	0.97

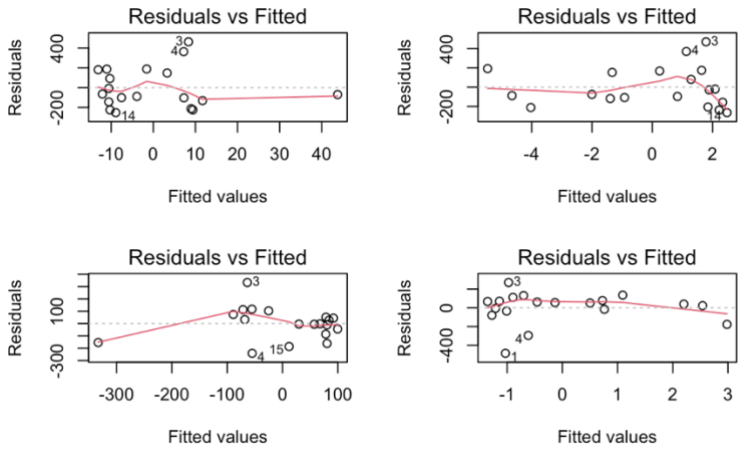

Figure S6: Residual vs. fitted plots for pairwise linear regression of PM 2.5 FPCA scores on wind speed FPCA scores.

The dataset and working code could be found at:

https://github.com/yixlu/FunctionalDataAnalysis_PM2.5.