Analyzing Effects of Meteorological Factors on PM 2.5 Concentration

Yichu Chen
University of California, Davis
Prof. Hans-Georg Müller

1 Introduction

PM2.5 refers to particulate matters with diameters of 2.5 micrometers or less. PM2.5 can come from multiple sources, such as coal combustion, industrial activities and automobile emissions. PM2.5 imposes a huge threat to the human body as it can easily penetrate into the repository system, jeopardize lung functions, cause inflammations and even promote cancer (Xu et al., 2021). Noticeably, the fluctuations of PM2.5 level is closely associated with meteorological factors. One previous study has confirmed a strong linear association between PM2.5 concentration and relative humidity, temperature, wind speed and pressure across major cities in China (Yang et al., 2017). Beijing, in particular, was known to have extremely poor air quality and high levels of PM2.5 concentration owing to its rapid development during the 2010s. In this project, we further investigate the hourly variation of PM2.5 concentration in Beijing. In particular, we explore a more flexible approach using functional data analysis techniques to model the hourly fluctuations of PM2.5 concentration as a smooth function and dynamically capture the effects of meteorological factors, including temperature, humidity and pressure, on PM2.5 levels on an hourly basis.

The dataset we use consists of hourly readings of PM2.5 concentration (microgram per cubic meter) from the US Embassy in Beijing and hourly meteorological measurements at Beijing Capital International Airport, both collected from January 1, 2010 to December 31, 2014. We use hourly PM2.5 concentration as response variable and hourly records of temperature (Celsius), Dew point (Celsius) and atmospheric pressure (hPa) as predictors. After data cleaning, our sample data encompasses a total of 1570 days with 24 hourly records in each day. The dataset comes from a larger study (Liang et al., 2015) and can be accessed at UC Irvine Machine Learning Repository. In this project, I mainly use the R "fdapace" package to complete my analysis (Zhou et al., 2022).

2 Methodology

2.1 Functional Principal Component Analysis

We have a sample of n longitudinal data recorded at N time grids ($X \in \mathbb{R}^{n \times N}$), which we assume are independent, observed realizations of n smooth random functions $X_i(t)(1 \leq i \leq n)$ in L^2 defined in a bounded and closed time interval $\mathbb{T}$. We assume that the sample data have

a common but unknown mean function $\mathbb{E}X(t) = \mu(t)$ as well as a smooth (auto)-covariance function $Cov(X(s), X(t)) = \Sigma(s,t)$ where $t, s \in \mathbb{T}$. Under mild conditions, the decomposition $\Sigma(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t)$ holds with uniform convergence. λ_k are eigenvalues in descending order and ϕ_k are the corresponding orthonormal eigenfunctions. $A_{ik} = \int_{\mathbb{T}} (X_i(t) - \mu(t)) \phi_k(t) dt$ is defined as the k^{th} functional principal component, or score. By Karhunen-Loeve expansion in $L^2([0, \mathbb{T}])$, we have $X_i(t) = \mu(t) + \sum_{k=1}^{\infty} A_{ik} \phi_k(t) + \epsilon_i(t)$ with $\mathbb{E}\epsilon_i(t) = 0$ and $Var(\epsilon_i(t)) = \sigma^2(t)$.

The decomposition facilitates dimension reduction, such that a desirable amount of explained variance can be attained using the first K functional principal components from the above expansion. That is, $\check{X}_i^K(t) = \mu(t) + \sum_{k=1}^{K} A_{ik} \phi_k(t)$. The corresponding fraction of total variance explained (FVE) is $FVE(K) = \sum_{k=1}^{K} \lambda_k / \sum_{k=1}^{\infty} \lambda_k$. In this project, we choose the number of eigenfunctions for each random function such that at least 95% of variance is explained. For any fixed K, FPCA provides a basis expansion that explains most of the variation in $X(t)$ in $L^2([0, \mathbb{T}])$. In other words, FPCA also allows us to learn the underlying dynamic of $X(t)$ based on eigenfunctions. Thus, we visualize the modes of variation based on the first K eigenfunctions. The k^{th} mode of variation is the set of functions given in the range $\mu(t) \pm \alpha \sqrt{\lambda_k} \phi_k(t), t \in \mathbb{T}, \alpha \in [-A, A]$. In our case, we use $A = 2$ (J.-L. Wang et al., 2016).

2.2 Functional Concurrent Regression

We have our predictors $X_r(t) (r \in \{1,2,3\})$ and response $Y(t)$, which we assume are smooth functions defined on a common time interval $\mathbb{T} = [0, 23]$. We also assume that $Y(t)$ depends only on $X_r(t)$ at t and not on $X_r(s)$, for all $s \neq t$. In addition, we assume that the relationship between $Y(t)$ and $X_r(t)$ are pointwise linear for any fixed $t \in \mathbb{T}$. The functional concurrent regression model (Şentürk and Müller, 2010) is defined as:

$$\mathbb{E}[Y(t) \,|\, X_1(t), X_2(t), X_3(t)] = \beta_0(t) + \sum_{r=1}^{3} \beta_r(t) X_r(t)$$

For estimation of mean and covariance functions, model coefficients, bootstrap confidence intervals and cross-validation, please refer to Appendix II.

3 Results

Pre-processing In total, 218 observations (days) that contain missing values in at least one variable are removed. By visual examination, we also found an abnormal measurement for PM2.5 at January 13, 2013, where the PM2.5 concentration experienced surged from less than 200 microgram per cubic meter at 12 AM to nearly 1000 microgram per cubic meter at 1 AM, which is likely due to false measurement. We therefore remove the observation. After pre-processing, we have 1570 observations fully observed at all hourly time grids (Figure 9).

Notation We denote PM2.5 concentration (microgram per cubic meter) as $Y(t)$; temperature (Celsius) as $X_1(t)$; humidity (Dew point in Celsius) as $X_2(t)$ and pressure (hPa) as $X_3(t)$. Their corresponding mean functions are $\mu_Y(t)$, $\mu_{X_1}(t)$, $\mu_{X_2}(t)$ and $\mu_{X_3}(t)$.

3.1 Trend

We first study the daily trends of temperature, humidity (dew point), pressure and PM2.5. We applied local linear smoothing and obtained the estimated mean functions as well as their 1st

derivatives. The results are shown in Figure 1. As the 3 meteorological factors are closely related to each other, we can easily observe some shared patterns. For instance, the mean functions for pressure and dew point have very similar shape: both curves started to decline continuously at roughly 10AM and reached daily minimum at around 3PM. Their derivative curves also share similar characteristics from visual inspection. Based on plot (a), we observe that the daily trend of temperature can be viewed as two phases: from 5AM to 3PM, temperature increases; from 3PM to 5AM on the next day, temperature decreases. The rate of increase in temperature reached daily maximum at 10AM; the rate of decrease in temperature is the highest at 8PM. From plot (d), we observe that the daily PM2.5 concentration decreases from 12AM to 3PM and bounces back after 3PM. Interestingly, the time at which PM2.5 concentration touches its daily minimum coincides with the time where temperature is at the highest and pressure and dew point are at the lowest.

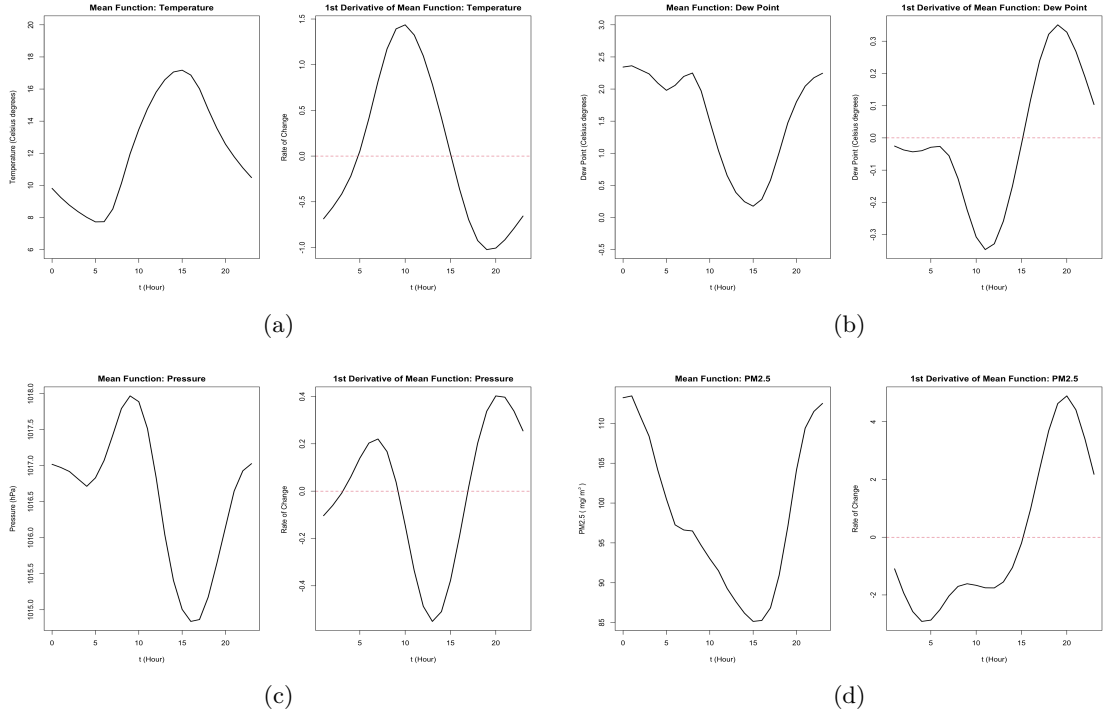


Figure 1: Mean Function and 1st Derivative

3.2 Variation

To further understand the general pattern of variation, we conduct functional principal component analysis (FPCA). Figure 4 in Appendix shows the 1st eigenfunction for temperature as well as the corresponding mode of variation. Based on our data, the first eigenfunction explains 96.97% of total variance in the temperature data. We notice that the first eigenfunction is nearly flat, which indicates that the day-to-day differences in average magnitude of temperature contributes to the largest variation. Similarly, both the 1st eigenfunctions of dew point and pressure (Figure 5 and 6 in Appendix) are nearly constant in time, indicating that the majority of variation is due to differences in daily averages, rather than hourly fluctuations around the mean functions. Specifically, the 1st eigenfunction of dew point explains 96.8% of variation in the data; the 1st eigenfunction of pressure explains 97.16% of variation.

The first 4 eigenfunctions for PM2.5 concentration and their modes of variation are shown in Figure 7 in Appendix. The 1st eigenfunction, which explains 73.31% of variation, is relatively constant in time, indicating that the majority of variation is due to differences in daily averages. The 2nd eigenfunction tells us that roughly 14.8% of variation in PM2.5 around the mean function are due to fluctuations before 10AM and after 12PM. The 3rd eigenfunction accounts for 5.97% of explained variance. Compared to the 2nd eigenfunction, the 3rd focuses more on variability near 10AM and less on variability near 5AM and 5PM. Since the 2nd eigenfunction accounts for over twice as much variance as the 3rd, we conclude that more variation around the mean curve occurs in the early morning and late afternoon than during the day. The 4th eigenfunction only explains 2.41% of variability left unexplained by the first 3.

3.3 Concurrent Regression Model

We fit the following regression model based on our data.

$$\mathbb{E}[Y(t)\,|\,X_1(t), X_2(t), X_3(t)] = \beta_0(t) + \sum_{r=1}^{3} \beta_r(t) X_r(t)$$

To begin with, we explored whether applying a lag to the predictors will give us better model performance. That is, we want to check the plausibility of predicting current PM2.5 concentration using past values of predictors. We adopted 5-fold cross validation with mean absolute error. The result is shown in Figure 8 in Appendix. It suggests that we should not apply any time lag to the predictors.

The time-varying coefficients and concurrent R^2 for our fitted model, along with their 95% bootstrap percentile confidence intervals with smoothed upper and lower bounds, are provided below. On average, the model is able to explain 28.54% of the total variance in PM2.5 concentration. The concurrent R^2 in Figure 2 suggests that our model has relatively good performance at around 9AM and 8PM, but weak performance at around 5AM.

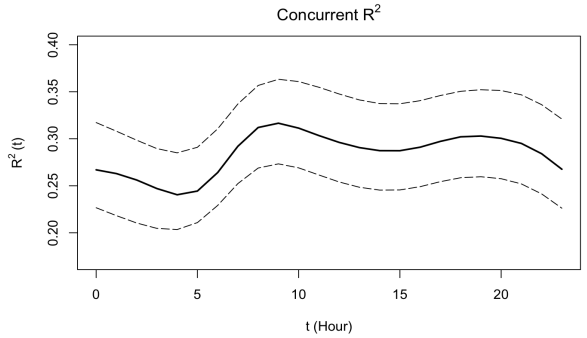
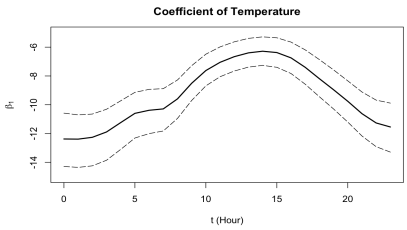
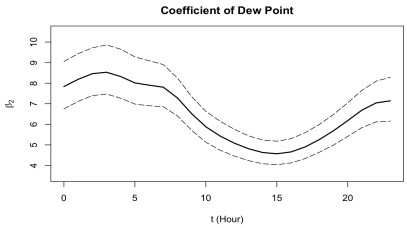


Figure 2: Concurrent R2

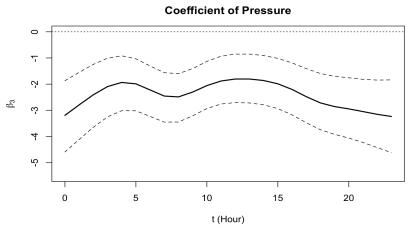
The 95% confidence intervals in Figure 3 suggest that all model coefficients are significantly nonzero at any time grid. Both temperature and pressure have negative association with PM2.5, while dew point has positive association. We also observed that the effect of temperature and dew point are the strongest from 12AM to 5AM and the weakest at roughly 3PM. The effect of pressure is slightly weaker between 8PM to 12AM and relatively stable otherwise.



(a)



(b)



(c)

Figure 3: Concurrent Regression Coefficients

4 Conclusion & Discussion

Our analysis concludes that all 3 meteorological factors have significant association with PM2.5 concentration at all time, from 12AM to 11PM. In particular, dew point is positively associated with PM2.5 levels; temperature and pressure are negatively associated with PM2.5 levels. Based on concurrent R^2, we conclude that the linear association between our meteorological predictors and PM2.5 concentrations are the strongest at around 8-9AM; the association is the weakest at around 5AM.

In section 3.1, we observed that there is close relationship between dew point and air pressure. The positive association between the two quantities can be justified, in a similar fashion of how increasing pressure also increases the boiling point of water ("Dew point," 2022). Our model suggested significant positive association between dew point and PM2.5 levels, which is consistent with the result from another similar study that concluded significant positive correlation between dew point and PM2.5 concentration (Xu et al., 2021). In fact, dew point can indirectly influence the level of PM2.5 pollutant. For instance, high dew point temperature promotes the formation of haze and fog, which can trap tiny particles in the air and cause PM2.5 to accumulate.

We also concluded that temperature and pressure are negatively associated with PM2.5. One possible explanation is that high relative temperature triggers vertical movement of air and creates convection currents that mix air at different altitudes, helping to disperse PM2.5. Still, some other studies conducted in different regions concluded positive correlation between temperature and PM2.5 concentration (J. Wang and Ogawa, 2015, for example). Whether or not such association is region-specific deserves further study. As for pressure, it has been established that lower pressure is linked with lower height of atmospheric boundary layer (ABL) and thus can lead to increased dispersion of air pollutants (Singh, 2019). On average, the height of ABL is the highest at roughly 3PM, which provides one possible explanation of why PM2.5 concentration is the lowest at the time.

In addition, we observed that the magnitude of effect of temperature and dew point on PM2.5 concentration continuously drops from 12AM to 3PM, and climbs up from 3PM to 12AM of the next day. The pattern is opposite to the general pattern of the mean function of PM2.5. In addition, our analysis based on eigenfunctions suggests that most of variability in temperature and dew point come from day-to-day differences in averages, and not hourly-specific variations. Combining the information, we suspect that there may be a natural decline in the effect of temperature and dew point when PM2.5 concentration decreases. More background knowledge may be needed to justify the validity of such hypothesis.

At last, it is worth mentioning that the dynamic of PM2.5 concentration can be highly volatile. In addition to temperature, dew point and pressure, other factors, such as sunlight exposure, wind speed, different weather conditions and even different geographical locations, can play important roles and change the pattern of variation in PM2.5. Thorough examination of additional likely factors will be required for a more accurate conclusion.

References

Dew point. (2022). https://en.wikipedia.org/wiki/Dew_point#cite_note-10

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., & Chen, S. X. (2015). Assessing beijing's pm2.5 pollution: Severity, weather impact, apec and winter heating. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 471(2182), 20150257. https://doi.org/10.1098/rspa.2015.0257

Şentürk, D., & Müller, H.-G. (2010). Functional varying coefficient models for longitudinal data. Journal of the American Statistical Association, 105(491), 1256–1264. https://doi.org/10.1198/jasa.2010.tm09228

Singh, A. (2019). Role of atmospheric boundary layer (abl) height and ventilation coefficient on urban air quality- a study based on observations and nwp model. Journal of Atmospheric Science Research, 2. https://doi.org/10.30564/jasr.v2i3.1421

Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2016). Functional data analysis. Annual Review of Statistics and Its Application, 3(1), 257–295. https://doi.org/10.1146/annurev-statistics-041715-033624

Wang, J., & Ogawa, S. (2015). Effects of meteorological conditions on pm2.5 concentrations in nagasaki, japan. International journal of environmental research and public health, 12, 9089–101. https://doi.org/10.3390/ijerph120809089

Xu, J., Liu, Z., Yin, L., Liu, Y., Tian, J., Gu, Y., Zheng, W., Yang, B., & Liu, S. (2021). Grey correlation analysis of haze impact factor pm2.5. Atmosphere, 2021, 1513. https://doi.org/10.3390/atmos12111513

Yang, Q., Yuan, Q., Li, T., Shen, H., & Zhang, L. (2017). The relationships between pm2.5 and meteorological factors in china: Seasonal and regional variations. International Journal of Environmental Research and Public Health, 14(12), 1510. https://doi.org/10.3390/ijerph14121510

Yao, F., Müller, H.-G., & Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association, 100(470), 577–590. https://doi.org/10.1198/016214504000001745

Zhou, Y., Bhattacharjee, S., Carroll, C., Chen, Y., Dai, X., Fan, J., Gajardo, A., Hadjipantelis, P. Z., Han, K., Ji, H., Zhu, C., Müller, H.-G., & Wang, J.-L. (2022). Fdapace: Functional

data analysis and empirical dynamics [R package version 0.5.9]. https://CRAN.R-project.org/package=fdapace

A Appendix I: Figures

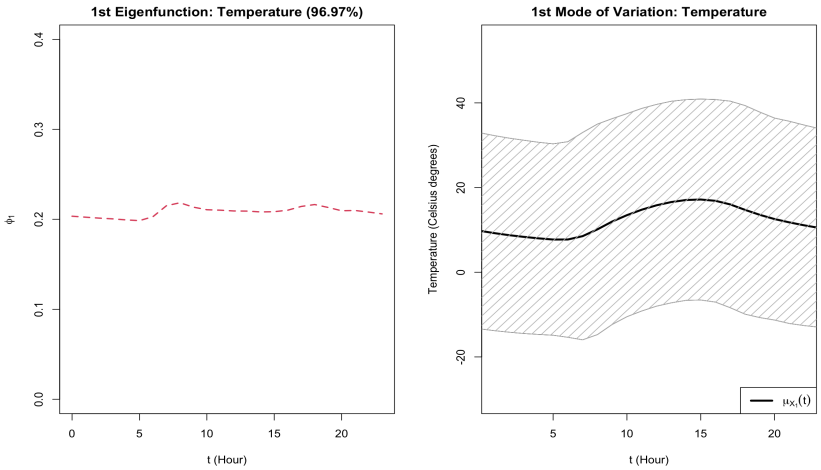


Figure 4: FPCA: Temperature

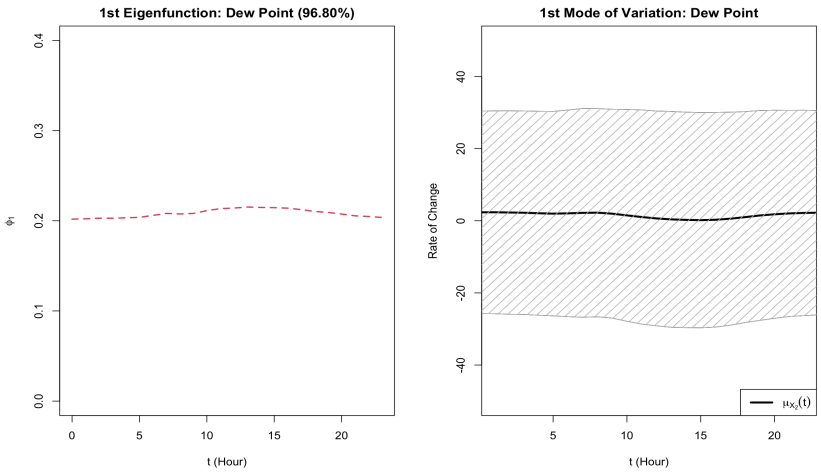


Figure 5: FPCA: Dew Point

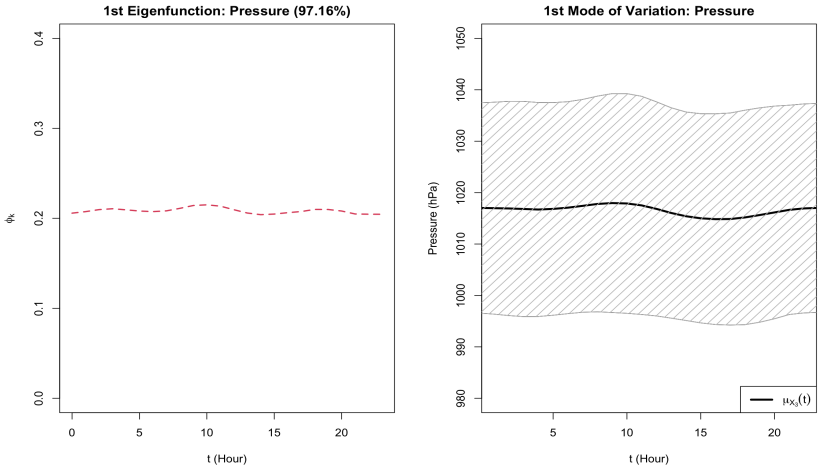
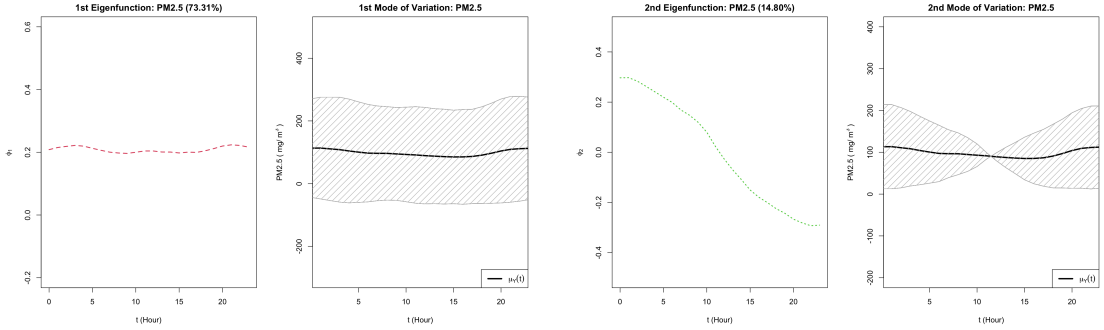
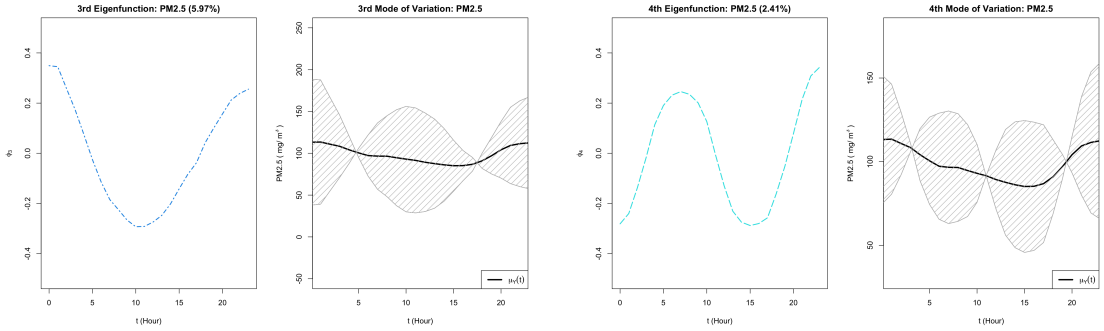


Figure 6: FPCA: Pressure



(a) (b)



(c) (d)

Figure 7: FPCA: PM2.5

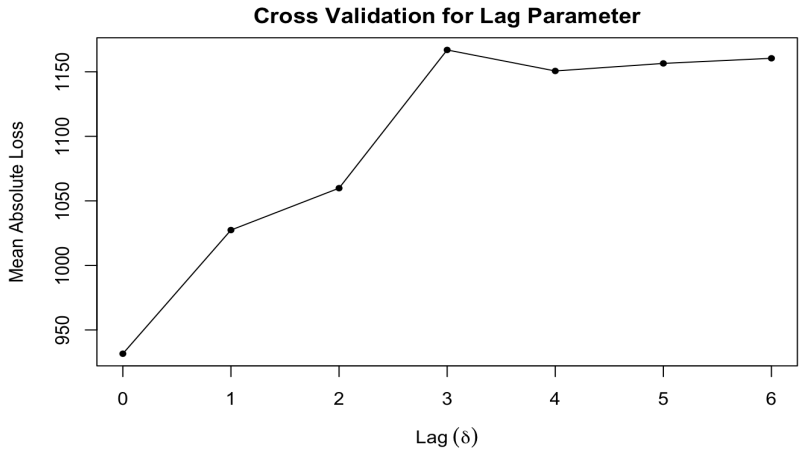


Figure 8: FPCA: Pressure

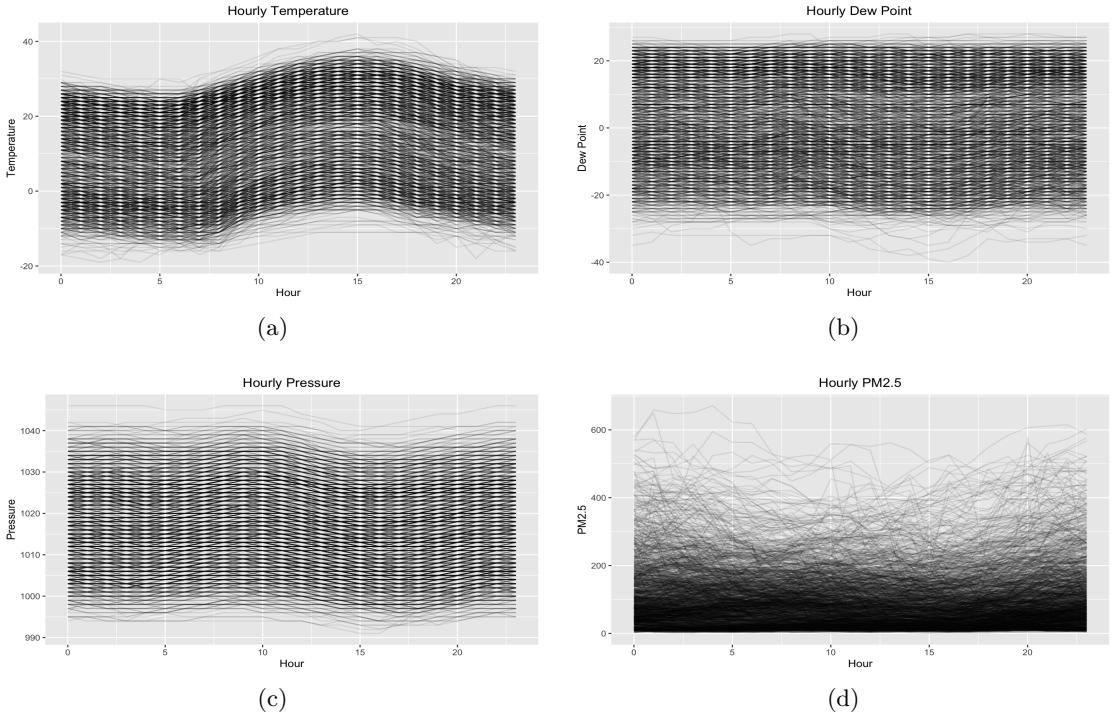


Figure 9: Distribution of Data Curves

B Appendix II: Methodology

B.1 Estimation of Mean and Covariance Function

We begin by assuming that the mean, covariance and eigenfunctions defined on $\mathbb{T} = [0, 23]$ are smooth; and that random error exists for all $t \in \mathbb{T}$. To estimate the mean function $\mathbb{E}X_i(t) = \mu(t)$, we use local linear scatterplot smoother and minimize (with respect to β_0, β_1) the following:

$$\sum_{i=1}^{N} \sum_{j=1}^{24} \{X_i(t_j) - \beta_0 - \beta_1(t - t_j)\}^2 K(\frac{t_j - t}{h_1})$$

The estimate of $\mu(t)$ is then $\hat{\mu}(t) = \hat{\beta}_0(t)$. Specifically, we use Gaussian kernel for the estimation. The tuning parameter h_1 is determined by generalized cross-validation (GCV).

To estimate the (auto)-covariance function $\Sigma_{XX}(s,t) = Cov(X(s), X(t))$ on $\mathbb{T} \times \mathbb{T}$, we apply two-dimensional scatterplot smoothing on raw covariances, which is defined as $C_{X_i X_i}(t_j, t_l) = (X_{ij} - \hat{\mu}(t_j))(X_{il} - \hat{\mu}(t_l))$, where $\hat{\mu}(t)$ is the estimated mean function. The diagonal elements were removed from the scatterplot before smoothing because they include additional variability due to random error. Using bivariate local linear smoothing, we minimize (with respect to $\beta_0, \beta_{11}, \beta_{12}$) the following:

$$\sum_{i=1}^{N} \sum_{1 \le j \ne l \le 24} \{C_{X_i X_i}(t_j, t_l) - \beta_0 - \beta_{11}(s - t_j) - \beta_{12}(t - t_l)\}^2 K(\tfrac{t_j - t}{h_2}) K(\tfrac{t_l - t}{h_2})$$

Using Gaussian kernel with bandwidth h_2 selected by GCV, the estimated covariance function is then $\hat{\Sigma}_{XX}(s,t) = \hat{\beta}_0(s,t),\ s \ne t$. Similarly, cross-covariances between $X(s)$ and $Y(t)$ can be estimated by 2D smoothing on the scatterplot of cross-covariances $C_{X_i Y_i}(t_j, t_l) = (X_{ij} - \hat{\mu}_X(t_j))(Y_{il} - \hat{\mu}_Y(t_l))$ on $\mathbb{T} \times \mathbb{T}$ based on estimated mean functions $\hat{\mu}_X(t)$ and $\hat{\mu}_Y(t)$ (Yao et al., 2005).

B.2 Estimation of Coefficients

Let $\hat{\Sigma}_{X_i X_j}(t,t)$ be the estimated covariance between $X_i(t)$ and $X_j(t)$ and $\hat{\Sigma}_{Y X_k}(t,t)$ the estimated covariance between $Y(t)$ and $X_k(t)$. Define $\hat{\chi}_t$ and $\hat{\Xi}_t$ as:

$$\hat{\chi}_t = \begin{bmatrix} \hat{\Sigma}_{X_1 X_1}(t,t) & \cdots & \hat{\Sigma}_{X_1 X_3}(t,t) \\ \vdots & \ddots & \vdots \\ \hat{\Sigma}_{X_3 X_1}(t,t) & \cdots & \hat{\Sigma}_{X_3 X_3}(t,t) \end{bmatrix} \quad \hat{\Xi}_t = \begin{bmatrix} \hat{\Sigma}_{Y X_1}(t,t) & \hat{\Sigma}_{Y X_2}(t,t) & \hat{\Sigma}_{Y X_3}(t,t) \end{bmatrix}^T$$

The coefficients can be estimated by $\begin{bmatrix} \hat{\beta}_1(t) & \hat{\beta}_2(t) & \hat{\beta}_3(t) \end{bmatrix}^T = \hat{\chi}_t^{-1} \hat{\Xi}_t$.

B.3 Bootstrap Confidence Interval

Under asymptotic consistency of estimated coefficients (cite!), we construct pointwise confidence intervals for the estimated coefficients at each time grid based on 1000 bootstrap samples generated with replacement. The bootstrap sample of estimates is $\{\hat{\beta}_{r,b}(t) : b = 1, ..., B\}$. A $(1-\alpha)100\%$ bootstrap percentile confidence interval for $\beta_r(t)$ is $[L_r, U_r]$, where L_r and U_r are the $[(B+1)\alpha/2]$-th and $[(B+1)(1-\alpha/2)]$-th smallest values in the bootstrap sample of estimates ($[x]$ is the nearest integer to x).

B.4 Cross Validation

We explored whether adding a lag to the predictor functions will improve model performance. We evaluated different choices of lag parameters δ ($\delta \ge 0$) using 5-fold cross validation with mean absolute error. Denote $Y_{i^*}(t)$ as the validation set ($i^* \in V$) and $\hat{Y}_{i^*,\delta}$ as the fitted value after applying time lag to predictors. The loss function at lag δ is defined as:

$$L(\delta) = \frac{1}{|V|} \sum_{i^* \in V} \sum_{j=1}^{24} |Y_{i^*}(t_j) - \hat{Y}_{i^*,\delta}(t_j)|$$