Prediction of Air Pollutants Concentrations by Nonparametric Regression Analysis

Jie Li, Qirui Hu

Center for Statistical Science and Department of Industrial Engineering,
Tsinghua University

2021 Symposium on Data Science and Statistics

Joint work with Qirui Hu

Introduction

- The environmental condition plays an important role in public health.
- Pollutants that have great impact on public health include carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), PM_{10}, $PM_{2.5}$ and sulfur dioxide (SO_2).
- The concentration of air pollutants will change rapidly within days.

Literature Review

- Air quality simulation: Tao et al. (2020) and Wang et al. (2014)
- The relationship between air pollutants: Zhang et al. (2015) and Wang et al. (2016)
- Air pollutants prediction: Wan et al. (2020) and Cabaneros et al. (2019)
- One serious drawback is that they only forecast air pollutants concentrations daily, failing to construct the prediction interval (PI).

Literature Review

- One common approach to deal with this problem is time series analysis, most based on parametric models such as autoregressive moving average (ARMA) model or autoregressive integrated moving average (ARIMA) model given the data is not stationary, see Mohd et al. (2009), Kumar and Jain (2010), Gocheva-Ilieva et al. (2014).

- Theoretical results on inference of the above models have been well developed, but the predicting performance varies in different scenarios.

- ARMA model or ARIMA model can be used for short-term prediction assuming that residuals are normal or asymptotically normal, but such assumption is relatively strict, which may cause that PI is always wider than usual in practice.

Literature Review

- A locally stationary model: $Y_t = m(t/T) + \sigma(t/T) Z_t$, see Dahlhaus (2012) and Dette and Wu (2020), only focused on the point predictor and ignored the PI.

- Little literature on the construction of prediction intervals of locally stationary time series; the only one is Das and Politis (2020), which novelly constructed both one-step-ahead point predictors and PIs for model-free or model-based scenario in the context of time series that are locally stationary.

- Wang et al. (2014) and Kong et al. (2018) successfully constructed multi-step-ahead PIs, but only focused on the simple AR(p) model.

Introduction

- In our paper, we introduce an novel and applicable method of forecasting multi-step-ahead future observation and establishing its corresponding PI under locally stationary model.
- The proposed method is applied to construct future air pollutants concentration PIs based on a quite large dataset that consists of 8 years' daily air pollutants data in Xi'an.

Data Description

- The data set consists of daily air pollutants concentrations in Xi'an between January 1, 2013 and July 31, 2020.
- Six major air pollutants concentrations: CO, NO_2, O_3, PM_{10}, $PM_{2.5}$ and SO_2, measured in ton per square kilometer (t/km^2), provided by the Xi'an Environmental Monitoring Center and CNEMC.
- The concentrations were obtained from the average of available hourly data measured at 13 state-controlled monitoring stations across Xi'an, in which one is the background site, the other 12 sites are involved in the calculation of pollutants concentration.

Data Description

Table 1: Descriptive statistics of each air pollutant concentration.

Pollutant	CO	NO_2	O_3	PM_{10}	$PM_{2.5}$	SO_2
Minimum	0.3	8	6	11	6	3
Maximum	5.7	129	301	903	589	163
Mean	1.38	47.40	96.31	120.77	65.65	20.53
Std. Dev.	0.73	18.37	56.39	85.39	60.59	20.56
Q1	0.9	33	50	65	29	8
Median	1.2	44	86	97	45	13
Q3	1.6	59	136	148.25	77	25
Skewness	1.71	0.73	0.65	2.49	2.78	2.79
Kurtosis	7.12	3.24	2.75	14.0	14.1	13.0

Data Description

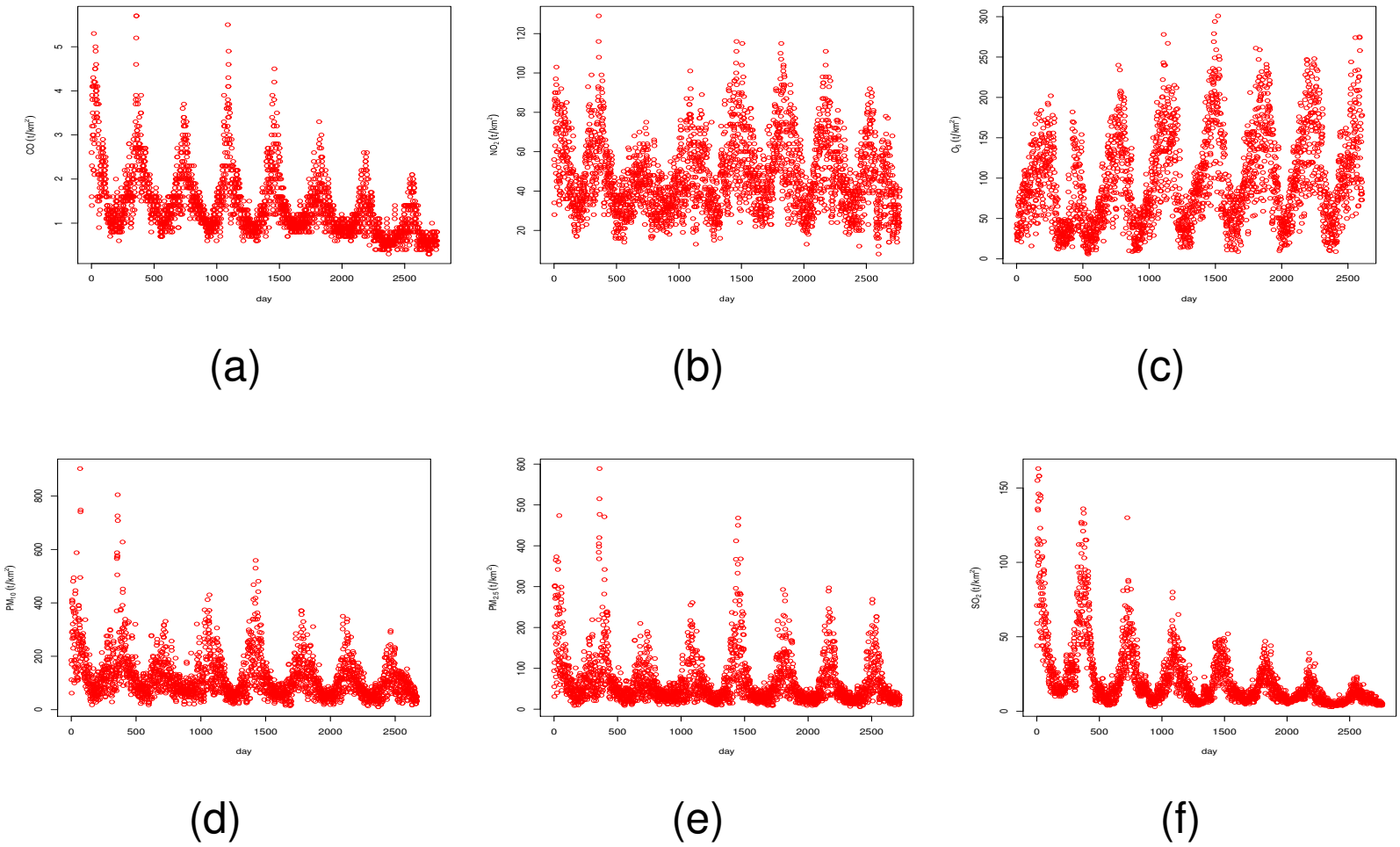

Figure 1: Scatterplot of each pollutant daily concentration.

Data Description

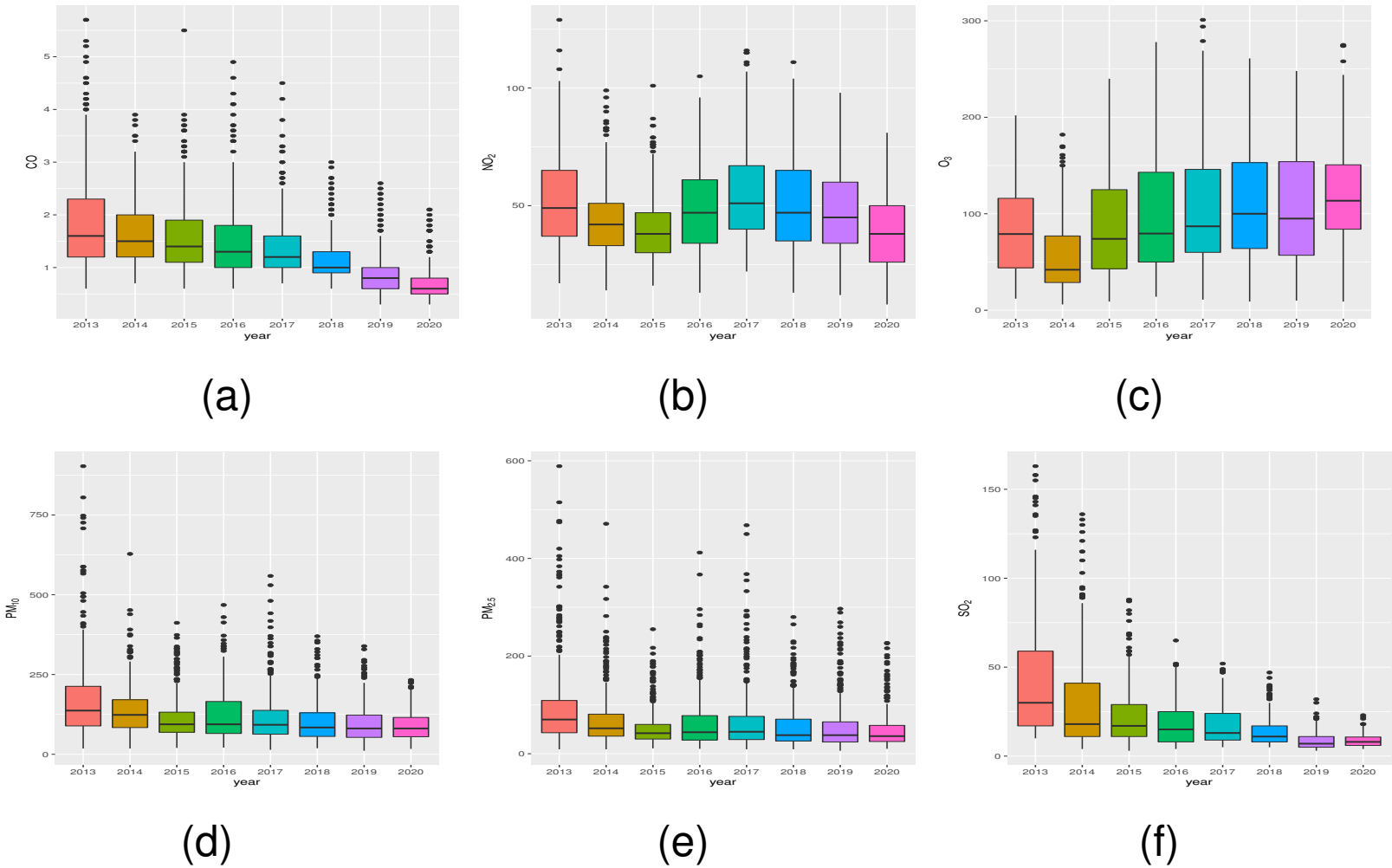

Figure 2: Box plot of each pollutant daily concentration. Black bars inside the boxes are the medians.

Nonparametric regression model

- One assumes that the observed time series $\{Y_t\}_{t=1}^{T}$ is a realization from the model

$$Y_t = m(t/T) + \sigma(t/T) Z_t, \quad t = 1, \cdots, T \tag{1}$$

- $m(\cdot)$ is the smooth time trend function, $\sigma^2(\cdot)$ is the conditional variance function.

- $\{Z_t\}_{t=1}^{T}$ is assumed to be stationary and weakly independent, satisfying $\mathbb{E}Z_t = 0$, $\mathbb{E}Z_t^2 = 1$, and is fitted the auto-regressive model with order p as follows:

$$Z_t = \sum_{k=1}^{p} \phi_k Z_{t-k} + \varepsilon_t, \quad t = p+1, \cdots, T, \tag{2}$$

where the white noise $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ is independent and identically distributed (IID) with mean 0.

Estimating the Trend Function $m(\cdot)$

- We propose to estimate the trend function $m(\cdot)$ by the following formula

$$\hat{m}(\cdot) = \underset{g(\cdot)\in\mathcal{H}^{(p-2)}}{\arg\min} \sum_{t=1}^{T} \{Y_t - g(t/T)\}^2 .$$

- The definition of $\hat{m}(\cdot)$ leads to

$$\hat{m}(\cdot) \equiv \sum_{J=1}^{J_s+p} \hat{\beta}_{J,p} B_{J,p}(\cdot),$$

- Applying the elementary algebra, one obtains

$$\hat{m}(\cdot) = \mathbf{B}(\cdot)^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

where the vector $\mathbf{Y}$ comprises noisy signals from raw data $\{Y_t\}_{t=1}^{T}$ with $\mathbf{Y} = (Y_1, \ldots, Y_T)^\top$.

Estimating the Variance Function $\sigma^2(\cdot)$

- Denote $e_t = Y_t - m(t/T)$, $t = 1, \ldots, T$, replace $m(\cdot)$ with the above estimator $\hat{m}(\cdot)$.
- The kernel estimator of $\sigma^2(\cdot)$ is

$$\hat{\sigma}^2(x) = \frac{\sum_{t=1}^{T} K_h(t/T - x)\hat{e}_t^2}{\sum_{t=1}^{T} K_h(t/T - x)}, \tag{3}$$

- $\hat{e}_t = Y_t - \hat{m}(t/N)$, $h = h_T > 0$ is the bandwidth and K is a kernel function with $K_h(u) = K(u/h)/h$.

Auto-regressive coefficient estimation

- The auto-regressive coefficients $\phi = (\phi_1, \cdots, \phi_p)^\top$ satisfy

$$\phi = \Gamma_p^{-1}\gamma_p, \quad \Gamma_p = \{\gamma(i-j)\}_{i,j=1}^{p}, \quad \gamma_p = (\gamma(1), \cdots, \gamma(p))$$

  in which $\gamma(l) = \mathbb{E}(Z_t Z_{t+l})$, $l = 0, \pm 1, \pm 2, \ldots$, represents the autocovariance function of $\{Z_t\}_{t=1}^{T}$.

- Denote $\widehat{Z}_t = \widehat{e}_t / \widehat{\sigma}(t/T)$ and the sample autocovariance function by

$$\widehat{\gamma}(l) = T^{-1} \sum_{t=1}^{T-l} \widehat{Z}_t \widehat{Z}_{t+l}, \quad 0 \le l \le T-1$$

- The classic Yule-Walker estimator of ϕ is a method of moment estimator based on the residuals $\{\widehat{Z}_t\}_{t=1}^{T}$ is defined by

$$\widehat{\phi} = \widehat{\Gamma}_p^{-1}\widehat{\gamma}_p, \quad \widehat{\Gamma}_p = \{\widehat{\gamma}(i-j)\}_{i,j=1}^{p}, \quad \widehat{\gamma}_p = (\widehat{\gamma}(1), \cdots, \widehat{\gamma}(p)) \qquad (4)$$

Constructing PI for Y_{T+k}

- The k-step-ahead linear predictor $\widetilde{Z}_{T+k}^{[k]}$ for Z_{T+k}, $k \geq 1$ based on $\{Z_t\}_{t=1}^{T}$ is defined recursively by

$$\widetilde{Z}_{T+k}^{[k]} = \phi_1 \widetilde{Z}_{T+k-1}^{[k-1]} + \cdots + \phi_p \widetilde{Z}_{T+k-p}^{[k-p]},$$

and satisfies

$$\widetilde{Z}_{T+k}^{[k]} = \phi_1^{[k]} Z_T + \cdots + \phi_p^{[k]} Z_{T-p+1},$$

in which the coefficient vector $\boldsymbol{\phi}^{[k]} = (\phi_1^{[k]}, \cdots, \phi_p^{[k]})^\top$ is a polynomial function g_k of $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_p)^\top$: $\boldsymbol{\phi}^{[k]} = g_k(\boldsymbol{\phi})$, with g_k defined by repeated applications of (2).

Constructing PI for Y_{T+k}

- With the Yule-Walker estimator $\widehat{\boldsymbol{\phi}} = (\widehat{\phi}_1, \cdots, \widehat{\phi}_p)^\top$ of $\boldsymbol{\phi}$ in (4), one then obtains a plug-in estimate $\widehat{\boldsymbol{\phi}}^{[k]} = (\widehat{\phi}_1^{[k]}, \cdots, \widehat{\phi}_p^{[k]})^\top = g_k(\widehat{\boldsymbol{\phi}})$ of $\widehat{\boldsymbol{\phi}}^{[k]} = g_k(\widehat{\boldsymbol{\phi}})$. Denote $\widehat{Z}_{T+k}^{[k]}$ as the data version of the linear predictor $\widetilde{Z}_{T+k}^{[k]}$:

$$\widehat{Z}_{T+k}^{[k]} = \widehat{\phi}_1^{[k]} \widehat{Z}_T + \cdots + \widehat{\phi}_p^{[k]} \widehat{Z}_{T-p+1},$$

and $\widehat{\varepsilon}_{T+k}^{[k]} = Z_{T+k} - \widehat{Z}_{T+k}^{[k]}$ as the k-step-ahead prediction residuals.

Constructing PI for Y_{T+k}

- Denote $F^{[k]}(x)$ the k-step-ahead prediction residual distribution and its α-th quantile as $q_\alpha^{[k]}$.

- $\hat{q}_{n,\alpha}^{[k]} = \left(\widehat{F}^{[k]}\right)^{-1}(\alpha) = \inf\left\{x : \widehat{F}^{[k]}(x) \geq \alpha\right\}$ of $q_\alpha^{[k]}$ based on a two-step plug-in kernel distribution estimator (KDE) $\widehat{F}^{[k]}(x)$ of $F^{[k]}(x)$

$$\widehat{F}^{[k]}(x) = \int_{-\infty}^{x} T^{-1} \sum_{t=k}^{T} \tilde{K}_{\tilde{h}}\left(u - \hat{\varepsilon}_t^{[k]}\right) du, \quad x \in \mathbb{R}, \tag{5}$$

where $\tilde{h} = \tilde{h}_T > 0$ is the bandwidth and $\tilde{K}$ is a kernel function with $\tilde{K}_{\tilde{h}}(u) = \tilde{K}\left(u/\tilde{h}\right)/\tilde{h}$, and $\hat{\varepsilon}_t^{[k]} = \widehat{Z}_t - \widehat{Z}_t^{[k]} = \widehat{Z}_t - \widehat{\phi}_1^{[k]}\widehat{Z}_{t-k} - \cdots - \widehat{\phi}_p^{[k]}\widehat{Z}_{t-k+p+1}$, $k \leq t \leq T$ are the prediction residuals.

Constructing PI for Y_{T+k}

- Combining the k-step-ahead predictor $\widehat{Z}_{n+k}^{[k]}$ and its corresponding quantile estimator, the $(1-\alpha)$-th PI for k-step-ahead observation Y_{T+k} is constructed as

$$\left[\widehat{m}(1)+\widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]}+\hat{q}_{n,\alpha/2}^{[k]}\right),\widehat{m}(1)+\widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]}+\hat{q}_{n,1-\alpha/2}^{[k]}\right)\right].$$

Implementation

- The spline estimator $\hat{m}(\cdot)$ with the number of interior knots $J_s = [6N^{1/4} \log\log N] + 1$, where $[a]$ denotes the integer part of a and spline order $p = 4$.

- For kernel estimator $\hat{\sigma}^2(x)$, choose the quartic kernel $K(u) = 15\left(1-u^2\right)^2 I\{|u| \leq 1\}/16$ and the bandwidth $h = 0.2h_{rot} \times \log^{-1/2} T$, where the rule-of-thumb bandwidth is

$$h_{rot} = \left[\frac{35 \sum_{t=1}^{T} \left\{ \widehat{e}_t^2 - \sum_{k=0}^{4} \widehat{a}_k (t/T)^k \right\}^2}{n \sum_{t=1}^{T} \left\{ 2\widehat{a}_2 + 6\widehat{a}_3 (t/T) + 12\widehat{a}_4 (t/T)^2 \right\}^2} \right]^{1/5},$$

in which
$\left(\widehat{a}_k\right)_{k=0}^{4} = \operatorname{argmin}_{\left(\widehat{a}_k\right)_{k=0}^{4} \in \mathbb{R}^5} \sum_{t=1}^{T} \left\{ Y_t - \sum_{k=0}^{4} a_k (t/n)^k \right\}^2$.

Implementation

- The estimation about AR time series (2) is carried out by using $\widehat{Z}_t$ in place of Z_t. The order p is determined by Akaike Information Criterion (AIC).

- After getting the Yule-Walker estimator of $\widehat{\boldsymbol{\phi}} = (\widehat{\phi}_1, \cdots, \widehat{\phi}_p)^\top$, we can obtain $\widehat{\boldsymbol{\phi}}^{[k]}$ by the following recursive formula:

$$\begin{aligned}
\widehat{\phi}_m^{[k]} &= \widehat{\phi}_1^{[k-1]}\widehat{\phi}_m + \widehat{\phi}_{m+1}^{[k-1]}, \quad 1 \le m \le p-1, \\
\widehat{\phi}_p^{[k]} &= \widehat{\phi}_1^{[k-1]}\widehat{\phi}_p,
\end{aligned}$$

with $\widehat{\phi}_m^{[0]} = \widehat{\phi}_m$ for $m = 1, \ldots, p$.

Implementation

- To estimate the quantile $q_{\alpha}^{[k]}$, the same kernel $\tilde{K}(u) = 15\left(1-u^2\right)^2 I\{|u| \leq 1\}/16$ is used in (5). The bandwidth h is taken to be $h = (4/3T)^{1/5}\hat{s}$, where $\hat{s}^2$ is the sample variance.

- Finally, the $(1-\alpha)$-th confidence prediction interval for future observations Y_{T+k} is constructed as

$$\left[\widehat{m}(1) + \widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]} + \hat{q}_{n,\alpha/2}^{[k]}\right), \widehat{m}(1) + \widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]} + \hat{q}_{n,1-\alpha/2}^{[k]}\right)\right].$$

Simulation studies

- The data is generated from the following model:

$$Y_t = m(t/T) + \sigma(t/N)\, Z_t, \quad t = 1, \cdots, N, \tag{6}$$

$$Z_t = 0.8 Z_{t-1} + \varepsilon_t, \quad t = 2, \cdots, N. \tag{7}$$

- $m(x) = 5 + 4\cos(2.5\pi x)$ and $\sigma(x) = (5 - \exp(-x))/(5 + \exp(-x))$ in (6). The iid errors $\{\varepsilon_t\}_{t=2}^{N}$ follow three different distributions: normal distribution $N(0, 0.5^2)$, mixture normal distribution $0.5N(-0.5, 0.5^2) + 0.5N(0.5, 0.5^2)$ and Laplace distribution $Laplace(0, \sqrt{2}/4)$, which can ensure $\mathbb{E}Z_t^2 = 1$.

- The sample size N is taken to be 1005, 2005, 4005, 8005, 16005, and realizations $\{Z_t\}_{t=-999}^{N}$ of size $N + 1000$ are generated from (7), with the first 1000 values deleted to guarantee strict stationarity of $\{Z_t\}_{t=1}^{N}$.

Simulation studies

- The dataset $\{Y_t\}_{t=1}^{N}$ is divided into a testing set $\{Y_t\}_{t=T+1}^{N}$ and a training set $\{Y_t\}_{t=1}^{T}$ with $T = N - 5$.

- 90% proposed PI:
$\left[\widehat{m}(1) + \widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]} + \hat{q}_{n,0.05}^{[k]}\right), \widehat{m}(1) + \widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]} + \hat{q}_{n,0.95}^{[k]}\right)\right]$
and 95% proposed PI:
$\left[\widehat{m}(1) + \widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]} + \hat{q}_{n,0.025}^{[k]}\right), \widehat{m}(1) + \widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]} + \hat{q}_{n,0.975}^{[k]}\right)\right]$.

- 90% normal PI:
$\left[\widehat{m}(1) + \widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]} - 1.64\hat{s}(k)\right), \widehat{m}(1) + \widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]} + 1.64\hat{s}(k)\right)\right]$
and 95% normal PI:
$\left[\widehat{m}(1) + \widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]} - 1.96\hat{s}(k)\right), \widehat{m}(1) + \widehat{\sigma}(1)\left(\widehat{Z}_{T+k}^{[k]} + 1.96\hat{s}(k)\right)\right]$
where $\hat{s}(k)$ is standard deviation of $\widehat{\varepsilon}_t^{[k]}$ based on the naive presumption that $F^{[k]}(x)$ is normal.

Table 2: The 95% and 90% PIs' average length (inside the parentheses) and coverage frequencies of future points over 1000 replications with normal distribution errors $\{\varepsilon_t\}_{t=2}^{N}$.

Point	T	95% Proposed PI	95% Normal PI	90% Proposed PI	90% Normal PI
Y_{T+1}	1000	0.840(1.715)	0.838(1.705)	0.774(1.433)	0.776(1.431)
	2000	0.880(1.831)	0.876(1.819)	0.824(1.529)	0.834(1.527)
	4000	0.896(1.886)	0.896(1.876)	0.836(1.577)	0.834(1.574)
	8000	0.912(1.915)	0.910(1.906)	0.850(1.602)	0.852(1.600)
	16000	0.944(1.974)	0.944(1.969)	0.904(1.654)	0.904(1.652)
Y_{T+2}	1000	0.800(2.036)	0.802(2.071)	0.734(1.736)	0.738(1.738)
	2000	0.854(2.248)	0.858(2.262)	0.786(1.902)	0.776(1.898)
	4000	0.890(2.360)	0.890(2.362)	0.844(1.986)	0.840(1.982)
	8000	0.902(2.421)	0.902(2.419)	0.850(2.034)	0.846(2.030)
	16000	0.944(2.512)	0.940(2.509)	0.912(2.108)	0.910(2.105)
Y_{T+3}	1000	0.750(2.153)	0.762(2.226)	0.672(1.858)	0.674(1.868)
	2000	0.846(2.429)	0.852(2.469)	0.750(2.074)	0.758(2.072)
	4000	0.898(2.587)	0.902(2.604)	0.842(2.190)	0.838(2.185)
	8000	0.900(2.678)	0.902(2.684)	0.842(2.257)	0.838(2.252)
	16000	0.946(2.793)	0.948(2.793)	0.884(2.348)	0.884(2.344)
Y_{T+5}	1000	0.682(2.209)	0.700(2.329)	0.612(1.930)	0.632(1.954)
	2000	0.776(2.554)	0.778(2.629)	0.728(2.202)	0.726(2.206)
	4000	0.884(2.769)	0.892(2.807)	0.820(2.361)	0.816(2.356)
	8000	0.906(2.903)	0.910(2.919)	0.834(2.456)	0.834(2.450)
	16000	0.926(3.046)	0.928(3.053)	0.882(2.568)	0.886(2.562)

Table 3: The 95% and 90% PIs' average length (inside the parentheses) and coverage frequencies of future points over 1000 replications with mixture normal distribution errors $\{\varepsilon_t\}_{t=2}^{N}$.

Point	T	95% Proposed PI	95% Normal PI	90% Proposed PI	90% Normal PI
Y_{T+1}	1000	0.802(2.357)	0.820(2.419)	0.734(2.002)	0.740(2.030)
	2000	0.866(2.451)	0.876(2.526)	0.822(2.090)	0.828(2.120)
	4000	0.902(2.566)	0.912(2.655)	0.846(2.195)	0.850(2.228)
	8000	0.916(2.619)	0.934(2.723)	0.866(2.248)	0.870(2.285)
	16000	0.954(2.681)	0.960(2.794)	0.908(2.306)	0.912(2.345)
Y_{T+2}	1000	0.798(2.855)	0.810(2.938)	0.744(2.452)	0.746(2.465)
	2000	0.822(3.076)	0.828(3.136)	0.742(2.629)	0.750(2.632)
	4000	0.876(3.292)	0.880(3.344)	0.822(2.807)	0.822(2.806)
	8000	0.914(3.406)	0.918(3.455)	0.844(2.902)	0.844(2.900)
	16000	0.920(3.509)	0.924(3.560)	0.882(2.990)	0.880(2.987)
Y_{T+3}	1000	0.734(3.031)	0.772(3.156)	0.656(2.627)	0.664(2.648)
	2000	0.790(3.338)	0.804(3.422)	0.738(2.867)	0.736(2.871)
	4000	0.866(3.628)	0.872(3.687)	0.822(3.098)	0.824(3.094)
	8000	0.928(3.790)	0.928(3.834)	0.854(3.224)	0.850(3.218)
	16000	0.936(3.923)	0.938(3.963)	0.886(3.332)	0.886(3.326)
Y_{T+5}	1000	0.698(3.116)	0.718(3.300)	0.640(2.733)	0.642(2.769)
	2000	0.754(3.516)	0.764(3.642)	0.706(3.045)	0.702(3.056)
	4000	0.846(3.895)	0.854(3.976)	0.784(3.341)	0.784(3.337)
	8000	0.900(4.118)	0.908(4.171)	0.842(3.509)	0.840(3.500)
	16000	0.938(4.293)	0.938(4.332)	0.854(3.644)	0.848(3.635)

Table 4: The 95% and 90% PIs' average length (inside the parentheses) and coverage frequencies of future points over 1000 replications with Laplace distribution errors $\{\varepsilon_t\}_{t=2}^{N}$.

Point	T	95% Proposed PI	95% Normal PI	90% Proposed PI	90% Normal PI
Y_{T+1}	1000	0.886(1.801)	0.868(1.654)	0.814(1.422)	0.802(1.388)
	2000	0.902(1.934)	0.882(1.772)	0.834(1.506)	0.832(1.487)
	4000	0.926(2.016)	0.912(1.849)	0.876(1.558)	0.874(1.552)
	8000	0.954(2.100)	0.934(1.933)	0.882(1.619)	0.886(1.622)
	16000	0.950(2.140)	0.932(1.973)	0.900(1.647)	0.902(1.655)
Y_{T+2}	1000	0.824(2.070)	0.812(2.017)	0.764(1.722)	0.758(1.693)
	2000	0.886(2.307)	0.872(2.207)	0.826(1.881)	0.816(1.852)
	4000	0.912(2.455)	0.898(2.331)	0.854(1.975)	0.854(1.956)
	8000	0.932(2.584)	0.918(2.453)	0.878(2.066)	0.878(2.059)
	16000	0.946(2.647)	0.932(2.514)	0.898(2.109)	0.898(2.110)
Y_{T+3}	1000	0.768(2.155)	0.768(2.172)	0.694(1.835)	0.690(1.823)
	2000	0.858(2.462)	0.848(2.413)	0.796(2.048)	0.790(2.025)
	4000	0.902(2.660)	0.894(2.573)	0.842(2.178)	0.844(2.159)
	8000	0.910(2.826)	0.898(2.723)	0.848(2.294)	0.850(2.285)
	16000	0.952(2.906)	0.94(2.799)	0.894(2.350)	0.882(2.349)
Y_{T+5}	1000	0.704(2.195)	0.710(2.279)	0.618(1.898)	0.618(1.912)
	2000	0.808(2.563)	0.806(2.575)	0.740(2.170)	0.738(2.161)
	4000	0.860(2.819)	0.858(2.777)	0.808(2.345)	0.806(2.331)
	8000	0.904(3.030)	0.892(2.962)	0.844(2.495)	0.844(2.486)
	16000	0.916(3.139)	0.900(3.060)	0.852(2.569)	0.852(2.568)

Real data analysis

- There are 2769 observations for each pollutant from January 1, 2013 to July 31, 2020. We remove the invalid observations which only accounts for quite a small proportion and use the remainder as our data set.

- The records of each pollutant concentration are split into a testing set (the last 5 observations) and a training set (other observations).

Forecasts from seasonal ARIMA model

- Given the pollutant observations $\{Y_t\}_{t=1}^{T}$, the seasonal ARIMA can be written as:

$$\phi_p(L)\,\Phi_P(L^s)\,\nabla^d\nabla_s^D Y_t = \theta_q(L)\,\Theta_Q(L^s)\,\varepsilon_t$$

where L the lag distance operator satisfying $L^d Y_t = Y_{t-d}$, ∇ is differencing operator satisfying $\nabla = 1 - L$ and $\nabla_s = 1 - L^s$, ε_t is the white noise satisfying $\varepsilon_t \sim WN(0, \sigma^2)$, ϕ and Φ include non-seasonal and seasonal autoregressive parameters, θ and Θ include non-seasonal and seasonal moving average parameters, with $\phi(L) = 1 - \sum_{k=1}^{p} \phi_k L^k$, $\Phi(L^s) = 1 - \sum_{k=1}^{P} \Phi_k (L^s)^k$, $\theta(L) = 1 - \sum_{k=1}^{q} \theta_k L^k$ and $\Theta(L^s) = 1 - \sum_{k=1}^{Q} \Theta_k (L^s)^k$.

Forecasts from seasonal ARIMA model

- Seasonal ARIMA model with lag $s = 365$ is applied to each air pollutant concentration. We specify the orders p, d, q, P, D, Q by Bayersian information criterion (BIC) and then estimate the coefficients by least square method.

- It is straightforward to find that seasonal ARIMA model may perform well in fitting, but usually with large predicting variance and much smaller lower prediction bound, sometimes even being a negative number.

Traditional time series model

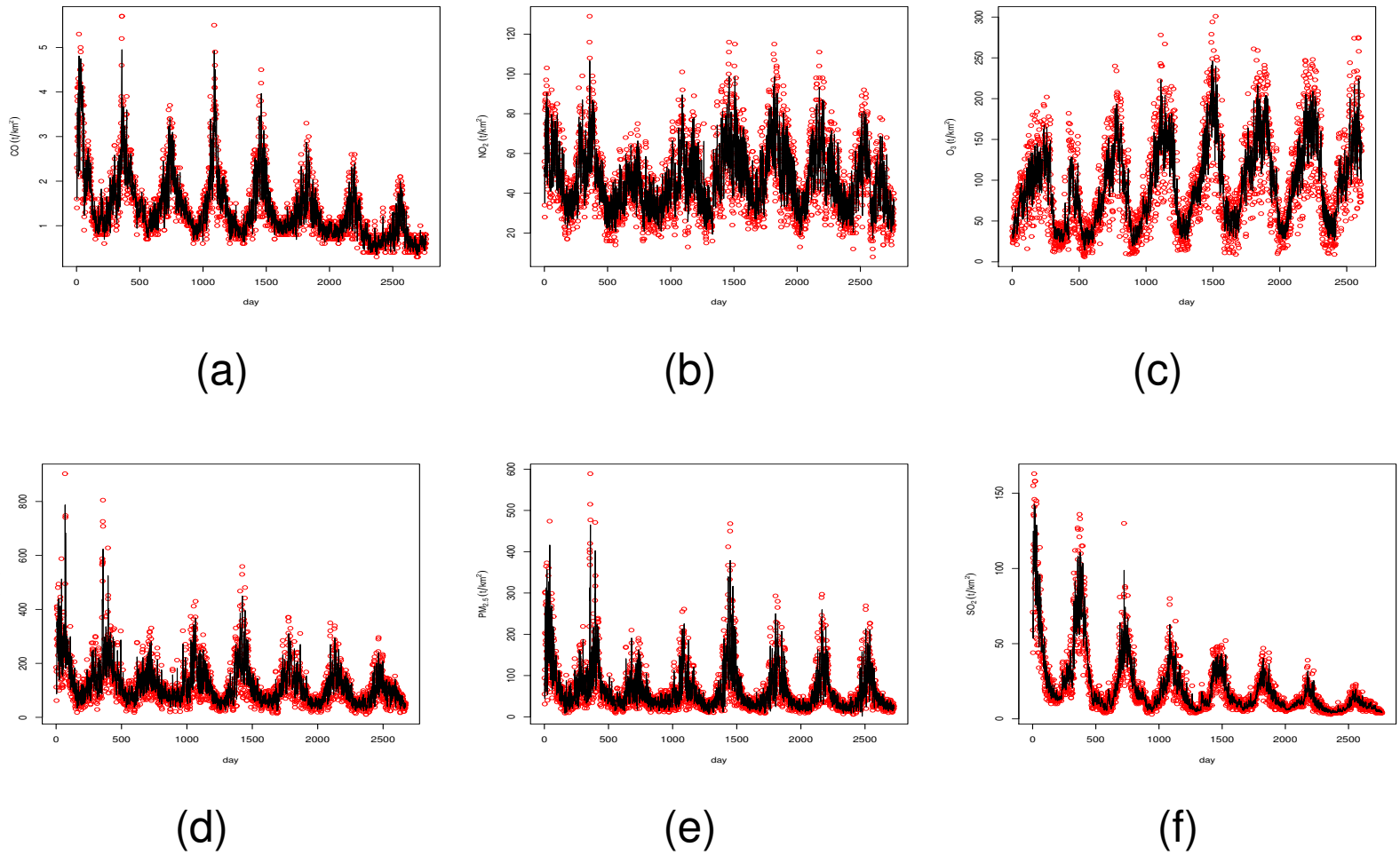

Figure 3: Fitted values of air pollutants concentrations with seasonal ARIMA model.

Table 5: 95% ARIMA PI and Proposed PI for future observations of each air pollutant.

Pollutant	Future observation	ARIMA PI	Proposed PI
CO	Y_{T+1}	$[0.0276, 1.211]$	$[0.555, 0.812]$
	Y_{T+2}	$[-0.057, 1.476]$	$[0.627, 0.888]$
	Y_{T+3}	$[-0.084, 1.533]$	$[0.632, 0.917]$
	Y_{T+4}	$[-0.107, 1.586]$	$[0.638, 0.950]$
	Y_{T+5}	$[-0.127, 1.635]$	$[0.606, 0.932]$
NO_2	Y_{T+1}	$[5.415, 48.356]$	$[16.881, 28.469]$
	Y_{T+2}	$[1.250, 55.604]$	$[16.312, 27.937]$
	Y_{T+3}	$[0.372, 57.816]$	$[15.745, 28.326]$
	Y_{T+4}	$[0.213, 58.584]$	$[15.286, 29.604]$
	Y_{T+5}	$[0.192, 59.012]$	$[16.085, 31.090]$
O_3	Y_{T+1}	$[70.028, 201.505]$	$[85.308, 190.703]$
	Y_{T+2}	$[47.000, 196.998]$	$[67.538, 174.426]$
	Y_{T+3}	$[36.341, 194.824]$	$[71.650, 188.302]$
	Y_{T+4}	$[32.698, 196.262]$	$[82.701, 205.975]$
	Y_{T+5}	$[38.401, 206.896]$	$[67.020, 195.309]$
PM_{10}	Y_{T+1}	$[-42.096, 152.525]$	$[36.663, 59.536]$
	Y_{T+2}	$[-69.120, 179.151]$	$[35.720, 58.666]$
	Y_{T+3}	$[-74.391, 184.265]$	$[33.157, 58.737]$
	Y_{T+4}	$[-74.739, 186.722]$	$[32.198, 60.000]$
	Y_{T+5}	$[-75.045, 189.103]$	$[32.410, 62.125]$
$PM_{2.5}$	Y_{T+1}	$[-43.433, 85.431]$	$[20.166, 33.182]$
	Y_{T+2}	$[-61.659, 110.086]$	$[24.184, 37.168]$
	Y_{T+3}	$[-64.626, 119.236]$	$[24.533, 38.780]$
	Y_{T+4}	$[-64.801, 121.879]$	$[23.427, 39.109]$
	Y_{T+5}	$[-65.069, 122.987]$	$[23.756, 40.127]$
SO_2	Y_{T+1}	$[-10.659, 22.008]$	$[4.049, 5.419]$
	Y_{T+2}	$[-13.846, 25.801]$	$[3.697, 5.068]$
	Y_{T+3}	$[-14.944, 27.342]$	$[3.446, 4.923]$
	Y_{T+4}	$[-15.506, 28.592]$	$[3.411, 5.052]$
	Y_{T+5}	$[-16.301, 30.210]$	$[3.472, 5.194]$

Nonparametric regression model

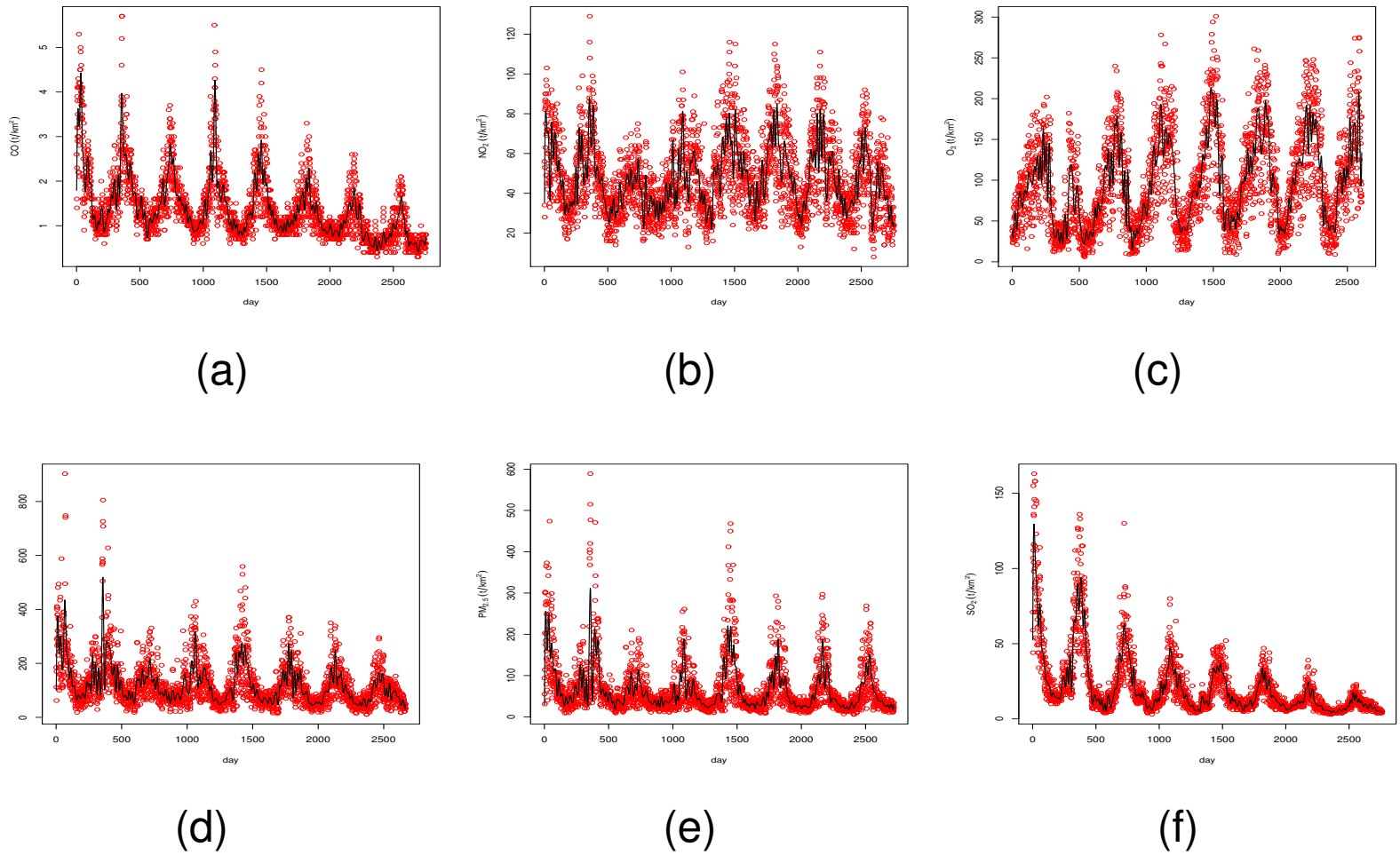

Figure 4: Each pollutant concentration with its trend function estimator $\hat{m}(\cdot)$ (solid line).

Forecasts by the proposed method

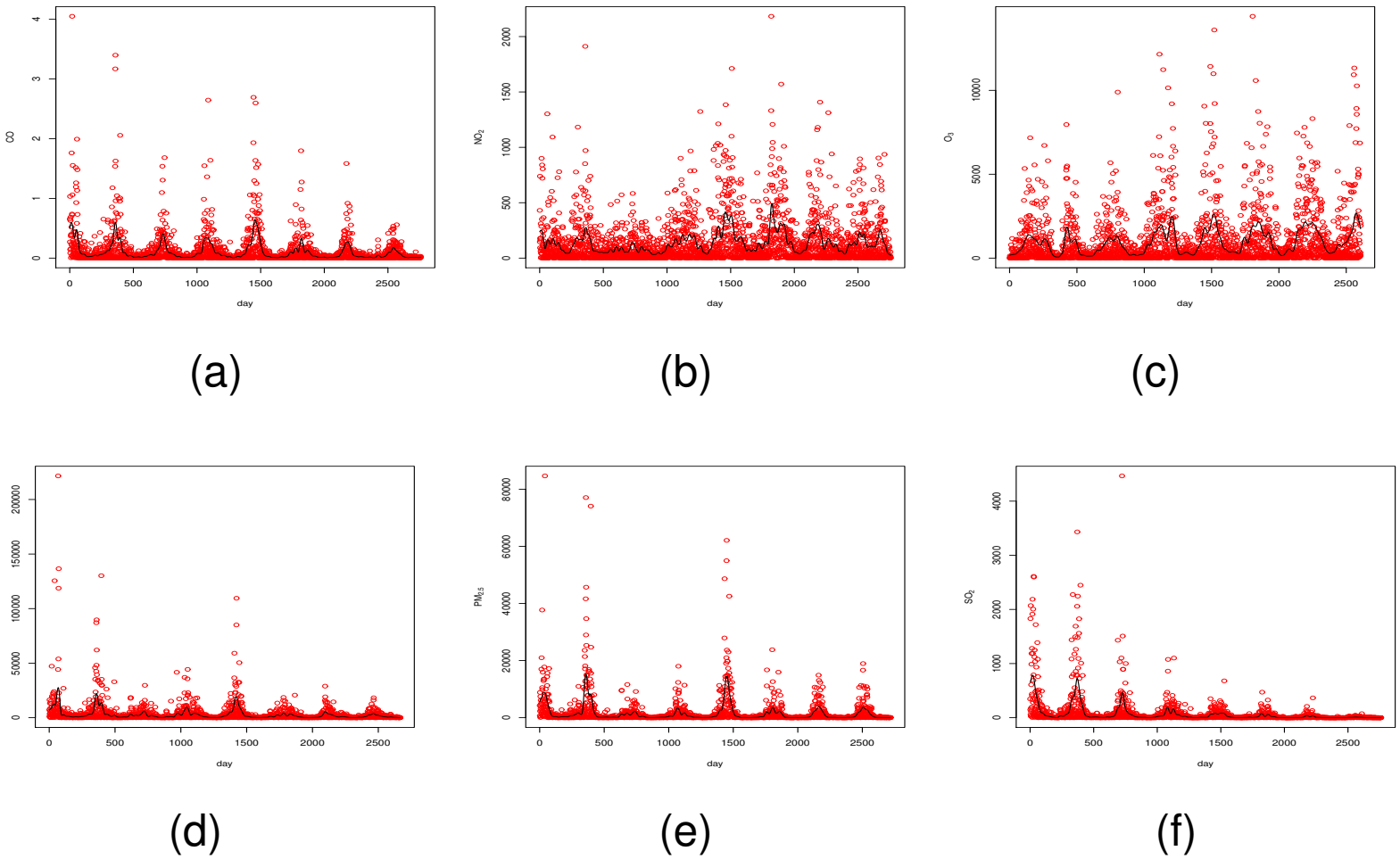

Figure 5: The scatterplot of $\hat{e}_t^2$ for each pollutant, its variance function estimator $\hat{\sigma}^2(\cdot)$ (solid line).

Forecasts by the proposed method

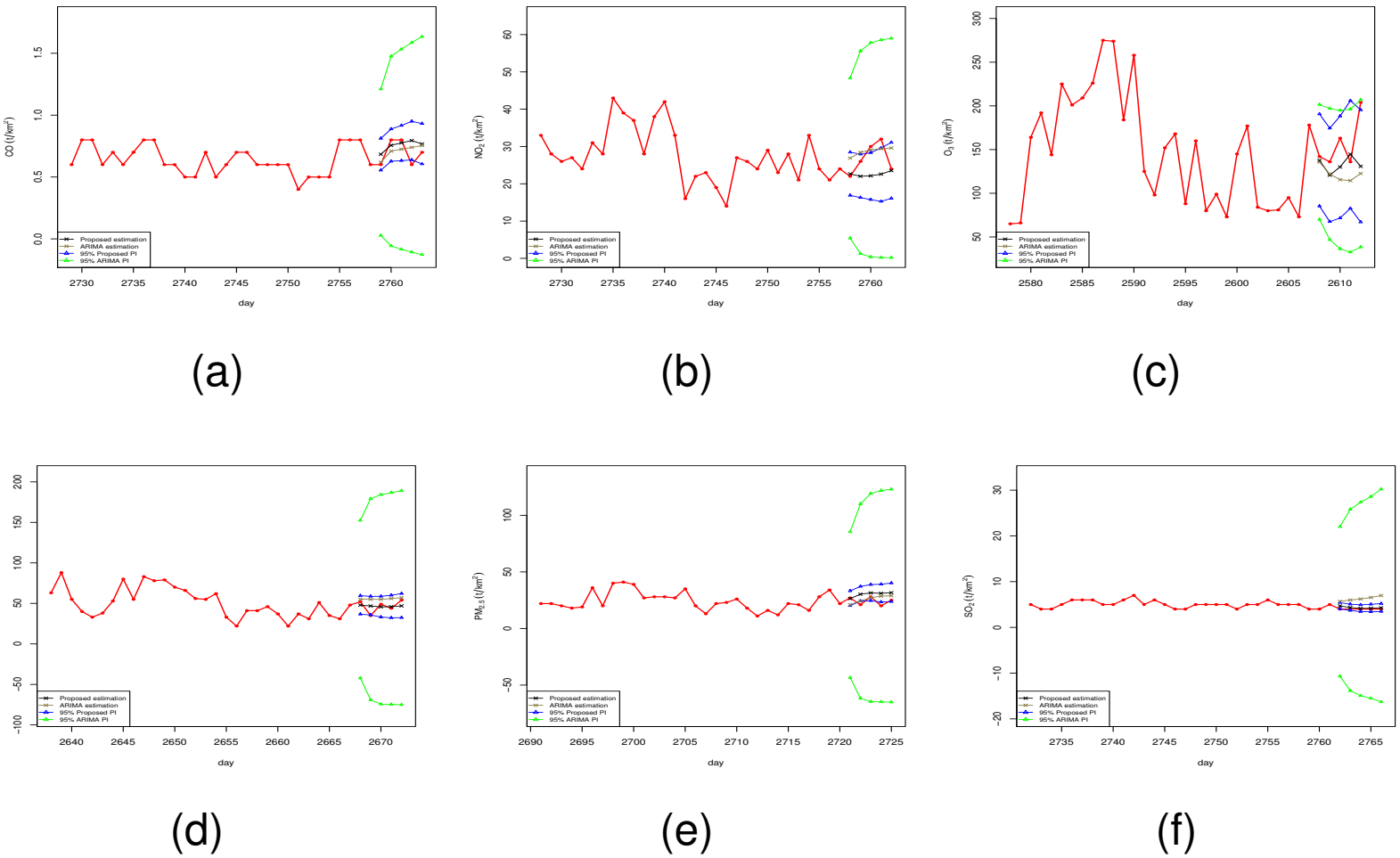

Figure 6: Plots of ARIMA estimation, Proposed estimation, 95% pointwise ARIMA PI and Proposed PI of the last 5 observations for each air pollutant concentration.

Concluding Remarks

- In our paper, we detect the global trend of each pollutant by applying nonparametric regression method. Our confidence intervals prove to be superior in comparison with those derived by traditional time series method.

- Extended study is expected to illustrate the mechanisms of $PM_{2.5}$ concentration with respect to corresponding factors, such as temperature, humidity and wind speed and to introduce temporal-spatial relationship into future study.

Acknowledgments

- Thank you for your attention!
- Any questions or comments?