Kriging Method for Missing Values in Spatial-temporal Data

Yuze Zhou

May.2022

Air-pollutant Data Description

- The air pollutant of concern is PM2.5, the data of which is collected from 9 monitoring sites in the city of Beijing.

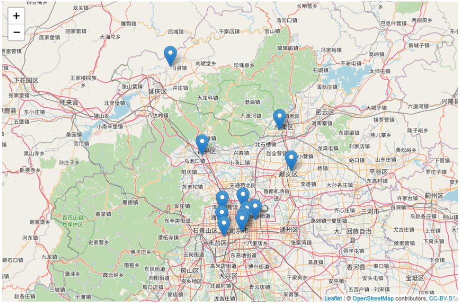


Figure 1: Location of Monitoring Sites

Air-pollutant Data Description

- For each site, the amount of PM2.5 is recorded hourly from Mar 1st 2013 to Dec 31 2013 with a total length of 7344 hours.
- Missing values for each site:

Aotizhongxin	Changping	Dingling	Dongsi	Guanyuan	Nongzhanguan	Tiantan
11	33	134	143	102	39	17

Wanliu	Wanshouxigong
21	41

Air-pollutant Data Description

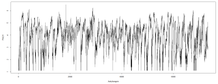
- The spatial-temporal process: $X(s;t)$, $s \in D_s$, $t \in D_t$
- D_s is the set of the location of all 9 sites, $D_t = \{1, \cdots, 7344\}$ is the collection of all time when the air-pollutant data is recorded.
- **Different levels in the mean value:**

Aotizhongxin	Changping	Dingling	Dongsi	Guanyuan	Nongzhanguan	Tiantan
82.392	72.645	64.755	86.940	82.129	84.348	83.200

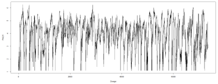
Wanliu	Wanshouxigong
91.596	84.028

Air-pollutant Data Description

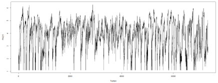
- Since the data is not normally distributed, transformation is applied to each of the process.



(a) Aotizhongxin



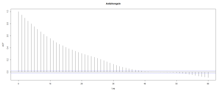
(b) Dongsi



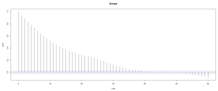
(c) Tiantan

Figure 2: Log-transformation

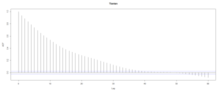
- **Stationarity:** The auto-correlation of process at each site suggests stationarity. Results from stationarity test like Dick-Fuller test also suggests it.



(a) Aotizhongxin



(b) Dongsi



(c) Tiantan

Figure 3: Autocorrelation Function

Spatial-temporal Process

- Denotes the process when applying log-transform on $X(s;t)$ as $Y(s;t)$. $Y(s;t)$ is stationary for each $s \in D_s$, but they have different levels in the mean value. We can assume

$$\mathbf{E}Y(s;t) := \mu(s_0, t_0) = \mu(s)$$

- Contour plot of the correlation function:

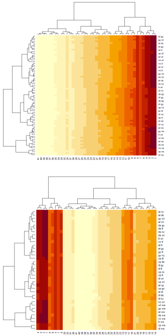


Figure 4: Contour Plot of Correlation

Spatial-temporal Process

- The stripe patterns suggest that the covariance function of $Y(s;t)$ is separable, which goes as followed:

$$cov(Y(\mathbf{s};t), Y(\mathbf{x};r)) = C^{(s)}(\mathbf{s};\mathbf{x})C^{(t)}(|t-r|)$$

- $C^{(s)}(\mathbf{s};\mathbf{x})$ is the spatial covariance function that only depends on the location of site s and x; $C^{(t)}(|t-r|)$ is the temporal covariance function that only depends on the time lag $|t-r|$
- Separability guarantees that the covariance function could be estimated from easily by obtaining the spatial covariance and the temporal covariance separably.

Spatial-temporal Kriging

- Assume we would like to predict the value of the process at site s_0 and time t_0 from observations $Z(s_i, t_{i,j}), i \in \{1, \cdots m\}, j \in \{1, \cdots, T_i\}$, the predictor $Y^*(s_0, t_0)$ is predicted using a linear combination of $Z(s_i, t_{i,j})$.

$$Y^*(s_0, t_0) = \sum_{i=1}^{m} \sum_{j=1}^{T_i} l_{ij} Z(s_i, t_{i,j}) + c$$

by denoting $l = \{l_{ij}\}$ and $\mathbf{Z} = \{Z_{ij}\}$, the predictor could be re-written as $Y^*(s_0, t_0) = l^{'} Z + c$, where l and c are the parameters to be estimated.

Spatial-temporal Kriging

- Since normality is satisfied for the transformed process, the conditional distribution of $Y(s_0, t_0)$ given $\mathbf{Z}$ is:
$$Y(s_0, t_0)|Z \sim \mathcal{N}(\mu(s_0, t_0) + c_0^{'} C_z^{-1}(\mathbf{Z} - \mu), c_{00} - c_0^{'} C_z^{-1} c_0)$$
where $C_z = var(\mathbf{Z})$, $c_0 = cov(Y(s_0, t_0), \mathbf{Z})$ and $c_{00} = var(Y(s_0, t_0))$.
- The simple kriging predictor $Y^*(s_0, t_0)$ is the mean of the conditional distribution:
$$Y^*(s_0, t_0) = \mathbf{E}(Y(s_0, t_0)|\mathbf{Z}) = \mu(s_0, t_0) + c_0^{'} C_z^{-1}(\mathbf{Z} - \mu)$$

Missing Values

- Assume missing value occurs at site s_0 at time t_0, we use the simple kriging method to predict $Y^*(s_0, t_0)$ as a replacement for the missing value.
- The contour plot and acf plot both suggest that the correlation would decay to near 0 after around 40 hours, therefore for observations before time $t_0 - 40$ and after time $t_0 + 40$ would not be very helpful in prediting $Y^*(s_0, t_0)$.
- The contour plot also suggests that the value of cross-correlation is high when the time lag is small for every pair of the sites; thus all sites $s \in D_s$ should be considered in the kriging predictor.

Missing Values

- The kriging predictor $Y^*(s_0, t_0)$ requires the usage of all existing observations $Z(j, t)$ such that $j \in D_s$ and $t \in [t_0 - 40, t_0 + 40]$.
- Kriging Predictor and actual observations

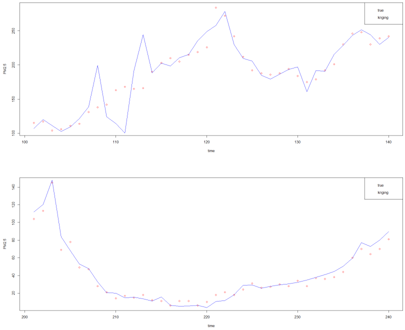


Figure 5: Kriging Predictor vs Actual Observations

- Kriging Predictor and missing values

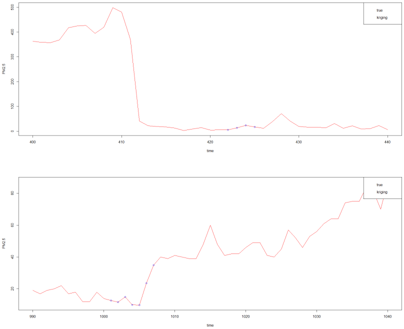


Figure 6: Kriging Predictor vs Actual Observations