Real-time PM2.5 mapping and anomaly detection from AirBoxes in Taiwan

G. Huang[1] L.-J. Chen[2] W.-H. Hwang[3] S. Tzeng[4] H.-C. Huang[5]

Introduction

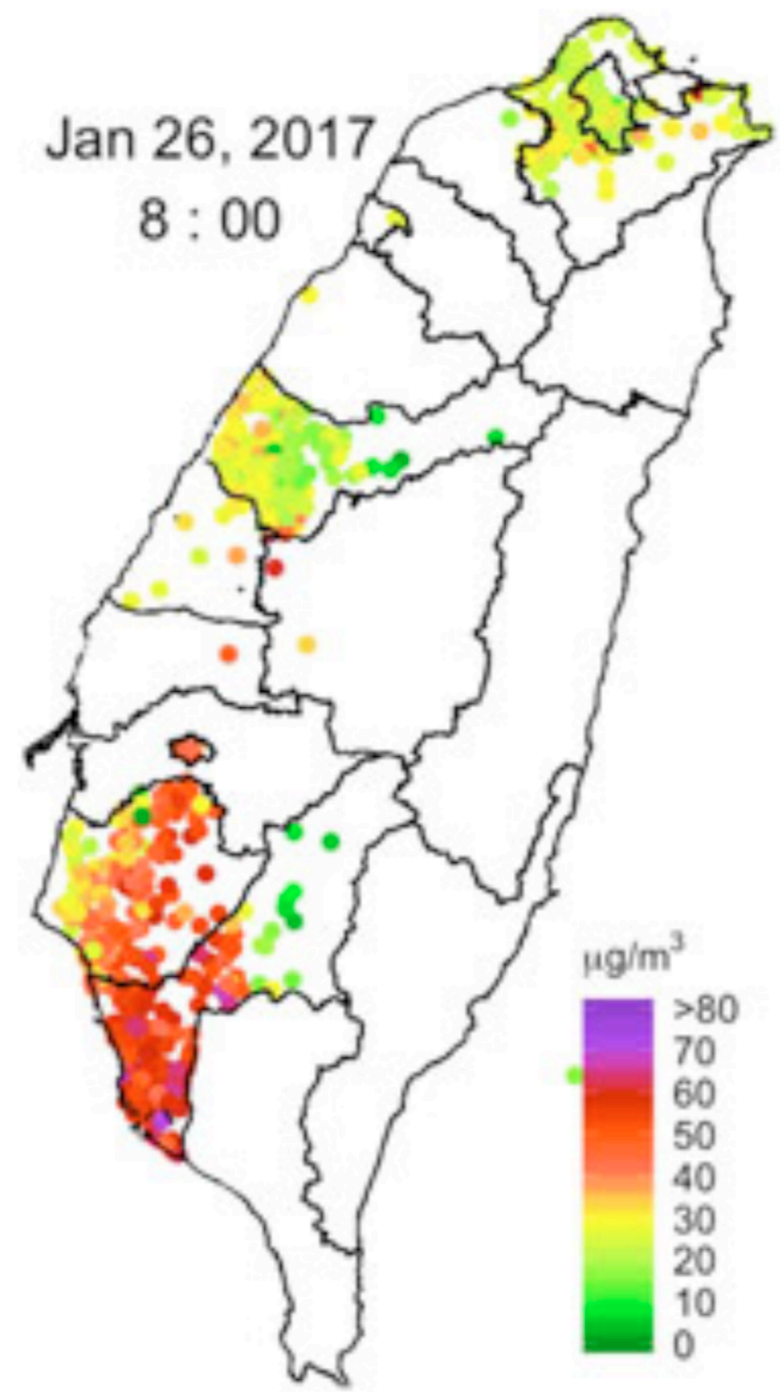
- What is PM2.5
- What is AirBox
- Anomaly detection
    - Clustering-based methods
    - Machine-learning methods
    - Statistical methods

Contribution

- Real time PM2.5 concentration at any location with its estimation error bar
    - Kriging approach
- A spatio-temporal control chart that can automatically monitor anomalous measurements by utilizing neighboring AirBox information.

AirBox Data

- 1283 AirBoxes across Taiwan from Jan. 1 to Feb. 28
- Precision is 111m * 102m
- Aggregate the data into hourly data at each location using the simple average
- Few unusual measurement that are either much higher or lower than nearby



Methodologies

regression coefficients

$$y(\boldsymbol{s}) = \boldsymbol{\beta}'\boldsymbol{\phi}(\boldsymbol{s}) + \eta(\boldsymbol{s}) = \sum_{k=1}^{K} \beta_k \phi_k(\boldsymbol{s}) + \eta(\boldsymbol{s}); \quad \boldsymbol{s} \in D,$$

first K multiresolution spline basis functions

zero-mean spatial process with an exponential covariance function

$$C(\boldsymbol{s}, \boldsymbol{u}) = \operatorname{cov}(y(\boldsymbol{s}), y(\boldsymbol{u})) = v_y^2 \exp(-\|\boldsymbol{s} - \boldsymbol{u}\|/\lambda)$$

- Suppose data $\boldsymbol{z} \equiv (z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n))'$
- With addictive white noise $\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{\varepsilon}$

where $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))'$, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, v_\varepsilon^2 \boldsymbol{I})$

Robust Method

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^K} \sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{\sigma}}\right)$$

$$e_i = z(\boldsymbol{s}_i) - \boldsymbol{\beta}' \boldsymbol{\phi}(\boldsymbol{s}_i)$$

- Where

$$\rho(x) = \begin{cases} x^2/2, & \text{if } |x| \le c \\ c|x| - c^2/2, & \text{if } |x| > c \end{cases}$$

Choose c = 1.345 which gives an efficiency of 95% if the error are normal distributed

$$\hat{y}(\boldsymbol{s}_0) = \hat{\boldsymbol{\beta}}' \boldsymbol{\phi}(\boldsymbol{s}_0) + \boldsymbol{c}(\boldsymbol{s}_0; \hat{\boldsymbol{\theta}})' \left(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) + \hat{v}_\epsilon^2 \boldsymbol{I}\right)^{-1} \left(\boldsymbol{z} - \left(\hat{\boldsymbol{\beta}}' \boldsymbol{\phi}(\boldsymbol{s}_1),\ \ldots,\ \hat{\boldsymbol{\beta}}' \boldsymbol{\phi}(\boldsymbol{s}_n)\right)'\right)$$

Anomaly Detection

- Baseline: $\{\hat{y}_t(\boldsymbol{s}) \, : \, \boldsymbol{s} \in D\}$
- Standardized residuals $r_t(\boldsymbol{s}_i) = \dfrac{z_t(\boldsymbol{s}_i) - \hat{y}_t(\boldsymbol{s}_i)}{\hat{\sigma}_t(\boldsymbol{s}_i)}; \quad i = 1, \ldots, n, \, t = 1, \ldots, T$

$r_t(\boldsymbol{s}_i)$ **has normal distribution if the parameter are known**

High positive r $\longrightarrow$ **It is higher than neighbor observation**

Low negative r $\longrightarrow$ **It is lower than neighbor observation**

do not need to specify a specific neighborhood range

Anomaly Detection

- Control chart of each AirBox: $r_t(\boldsymbol{s}_i); i = 1, \ldots, T$
- Control limits: 3 standard deviation: $|r_t(\boldsymbol{s}_i)| > 3$
- Ranking: $\mathrm{RMSE}_i = \left\{ \frac{1}{|T_i|} \sum_{t \in T_i} (r_t(\boldsymbol{s}_i))^2 \right\}^{1/2}$
- AirBoxes with high RMSE indicate that they tend to produce outlying observations

Decompose RMSE

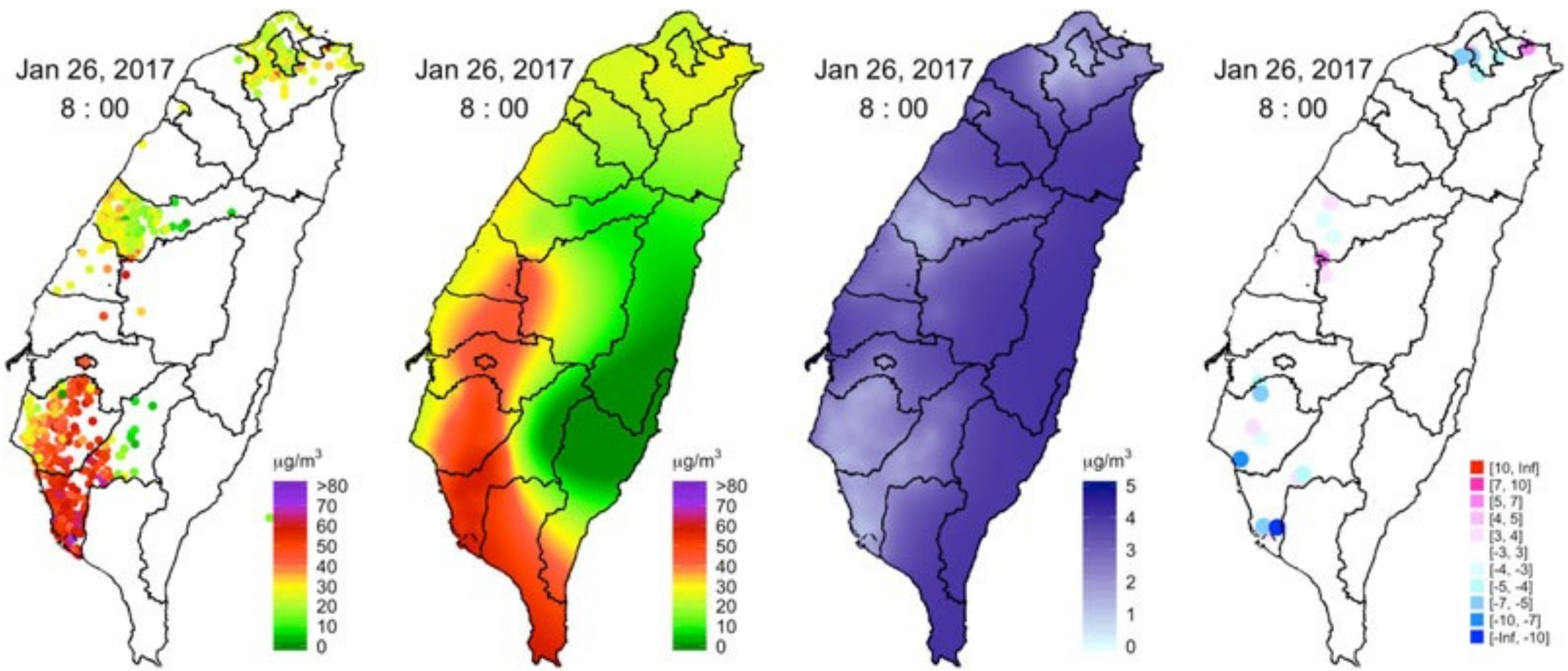
$$
\begin{aligned}
\mathrm{RMSE}_i &= \left(b_i^2 + V_i\right)^{1/2}, \\
b_i &= \frac{1}{|T_i|} \sum_{t \in T_i} r_t(\boldsymbol{s}_i), \\
V_i &= \frac{1}{|T_i|} \sum_{t \in T_i} \left\{ r_t(\boldsymbol{s}_i) - \frac{1}{|T_i|} \sum_{t \in T_i} r_t(\boldsymbol{s}_i) \right\}^2
\end{aligned}
$$

- Classification

	High RMSE
Bi <= -3	Average standardize residual is above control limit
Bi >= 3	Average standardize residual falls below control limit
-3 < Bi < 3	Tends to have high variation

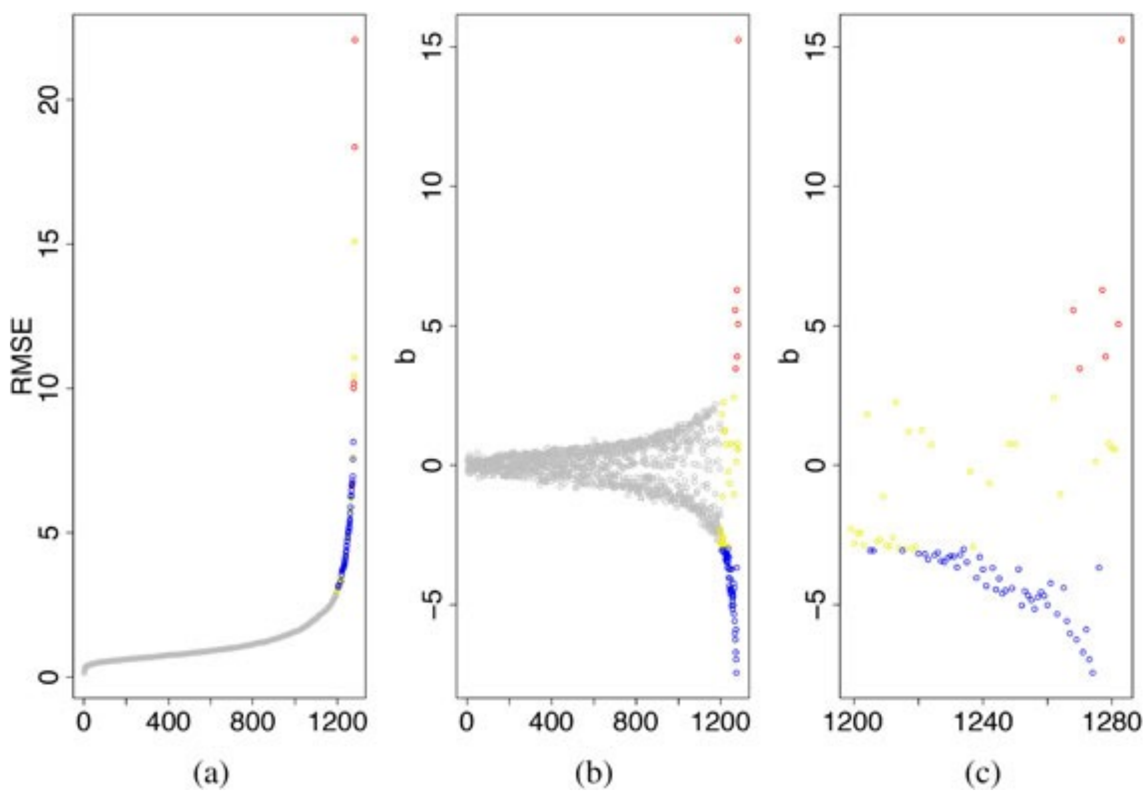
Analysis Result

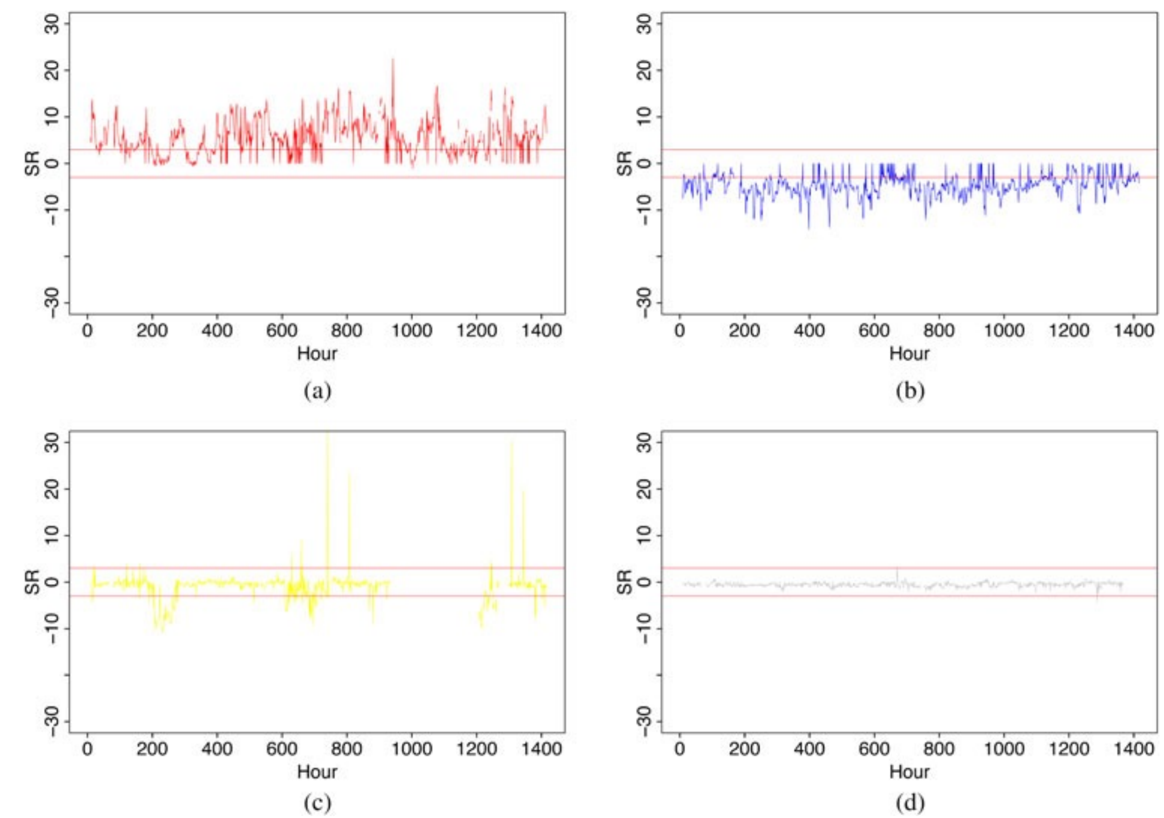
- Locations with fewer AirBox has higher error values
- Unusual large or small PM2.5 value is shown in fig. D



(a) (b) (c) (d)

Anomaly Detection Result





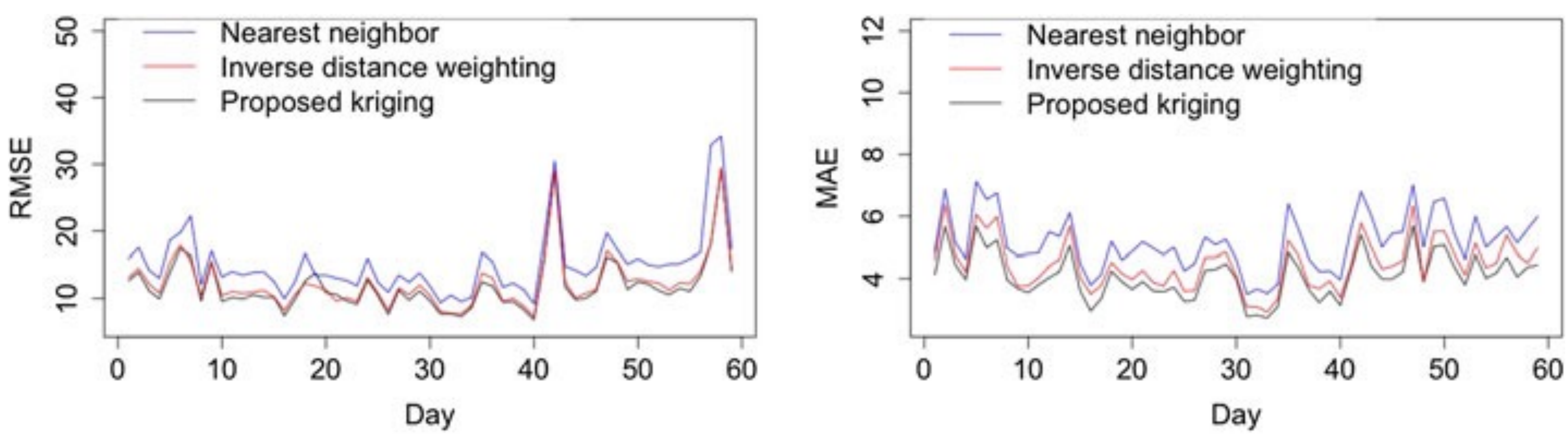
- (a) Ranking RMSE value
- (b) The corresponding biases of (a)
- (c) Classify high RMSE into 3 groups

Compare with Other Method

- Criteria

$$\mathrm{RMSE} = \left\{ \frac{1}{100} \sum_{i=1}^{100} \left(\tilde{y}\left(\boldsymbol{s}_i^*\right) - z\left(\boldsymbol{s}_i^*\right) \right)^2 \right\}^{1/2},$$

$$\mathrm{MAE} = \mathrm{median} \left\{ \left| \tilde{y}\left(\boldsymbol{s}_1^*\right) - z\left(\boldsymbol{s}_1^*\right) \right|, \ldots, \left| \tilde{y}\left(\boldsymbol{s}_{100}^*\right) - z\left(\boldsymbol{s}_{100}^*\right) \right| \right\}$$



Method	Averaged RMSE	Averaged MAE
Nearest neighbor	15.16	5.18
Inverse distance weighting	12.37	4.46
The proposed kriging	11.88	4.09

Conclusion

- Proposed method is able to detect potential emission sources, malfunctioned AirBoxes, and AirBoxes that are put indoors
- AirBoxes provides very high spatial and temporal coverage but they have much more higher error than those large monitoring stations