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

G. Huang¹ L.-J. Chen² W.-H. Hwang³ S. Tzeng⁴ H.-C. Huang⁵
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

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

Suppose data \( z \equiv (z(s_1), \ldots, z(s_n))' \)

With addictive white noise \( z = y + \epsilon \)

where \( y = (y(s_1), \ldots, y(s_n))' \), \( \epsilon \sim N(0, \sigma^2 \epsilon I) \)

\( C(s, u) = \text{cov}(y(s), y(u)) = \sigma^2 \exp(-\|s - u\|/\lambda) \)
Robust Method

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

\[ e_i = z(s_i) - \beta' \phi(s_i) \]

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

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

\[ \hat{y}(s_0) = \hat{\beta}' \phi(s_0) + c(s_0; \hat{\theta})' \left( \Sigma(\hat{\theta}) + \hat{v}_\epsilon^2 I \right)^{-1} \left( z - \left( \hat{\beta}' \phi(s_1), \ldots, \hat{\beta}' \phi(s_n) \right) \right) \]
Anomaly Detection

• Baseline: \( \{ \hat{y}_t(s) : s \in D \} \)

• Standardized residuals
  \[ r_t(s_i) = \frac{z_t(s_i) - \hat{y}_t(s_i)}{\hat{\sigma}_t(s_i)} ; \quad i = 1, \ldots, n, t = 1, \ldots, T \]

\[ r_t(s_i) \] has normal distribution if the parameter are known

High positive \( r \) \( \quad \rightarrow \) It is higher than neighbor observation

Low negative \( r \) \( \quad \rightarrow \) It is lower than neighbor observation

do not need to specify a specific neighborhood range
Anomaly Detection

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

\[ \text{RMSE}_i = \left( b_i^2 + V_i \right)^{1/2} , \]

\[ b_i = \frac{1}{|T_i|} \sum_{t \in T_i} r_t(s_i) , \]

\[ V_i = \frac{1}{|T_i|} \sum_{t \in T_i} \left\{ r_t(s_i) - \frac{1}{|T_i|} \sum_{t \in T_i} r_t(s_i) \right\}^2 \]

- **Classification**

<table>
<thead>
<tr>
<th><strong>High RMSE</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bi &lt;= -3</strong></td>
<td>Average standardize residual is above control limit</td>
</tr>
<tr>
<td><strong>Bi &gt;= 3</strong></td>
<td>Average standardize residual falls below control limit</td>
</tr>
<tr>
<td><strong>-3 &lt; Bi &lt; 3</strong></td>
<td>Tends to have high variation</td>
</tr>
</tbody>
</table>
Analysis Result

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

\[
\text{RMSE} = \left\{ \frac{1}{100} \sum_{i=1}^{100} (\tilde{y}(s_i^*) - z(s_i^*))^2 \right\}^{1/2},
\]

\[
\text{MAE} = \text{median} \left\{ |\tilde{y}(s_1^*) - z(s_1^*)|, \ldots, |\tilde{y}(s_{100}^*) - z(s_{100}^*)| \right\}
\]

<table>
<thead>
<tr>
<th>Method</th>
<th>Averaged RMSE</th>
<th>Averaged MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest neighbor</td>
<td>15.16</td>
<td>5.18</td>
</tr>
<tr>
<td>Inverse distance weighting</td>
<td>12.37</td>
<td>4.46</td>
</tr>
<tr>
<td>The proposed kriging</td>
<td>11.88</td>
<td>4.09</td>
</tr>
</tbody>
</table>
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