OPTIMIZED MONITOR REDUCTION FOR AN INDUSTRIAL AIR POLLUTION MONITORING NETWORK

M.D. ADAMS and P.S. KANAROGLOU

1School of Geography & Earth Sciences, McMaster University, 1280 Main Street West, Hamilton, ON, L8S 4L8, Canada.
adamsmd@mcmaster.ca

EXTENDED ABSTRACT

The Hamilton Air Monitoring Network (HAMN), which is located in Hamilton, Ontario, Canada, consists of 14 air quality monitors, 8 of which monitor particulate matter 10 microns or smaller in aerodynamic diameter. Operating these monitoring devices comes at a large expense, which is currently funded by industry partners located in Hamilton. However, long-term partner commitment is not currently insured and partners have withdrawn their support in the past. The objective of this network is to monitor industrial emissions, particularly to identify when air pollution levels are above prescribed thresholds. For HAMN the majority of funds to support the network are dedicated to annual operating costs. Thus, this monitoring network is very dependent on the sustained funding from the industrial partners. With Hamilton's declining traditional industrial sector a potential for partners to revoke funding is present, which has previously occurred.

The HAMN monitoring network is dense; with 14 monitors located within an 8 x 8 km square. Because of this high density some monitors may be redundant. We evaluate the scenario of reduced monitor operation to identify which monitors, if removed from the network, would result in the minimized loss of information on air pollution threshold exceedance. A combinatorial problem was formulated to select the sub-set of monitors from the full set to maximize the retention of exceedance air quality data obtained across the domain. The dataset used was the historical monitoring records between December 1st, 2010 and November 31, 2012. The spatial domain was divided into a discrete space of grid cells. The threshold concentration is a 24-hour average of 50 µg m⁻³. For each hour in the time-series, a pollution surface was interpolated with the full set of monitoring data for that hour. Using the previous 23 hours' concentrations the 24-hour average was determined for each grid cell. Each grid cell was assigned as above or below the threshold concentration. This process resulted in a set of grid cells each with a time-series of indicator values for threshold exceedance, and was the baseline data. To evaluate each reduced monitoring network configuration, we interpolated a surface with the data from the reduced monitor set. The 24-hour concentrations were determined. At each grid cell it was evaluated if the subset of monitors’ data correctly predicted the occurrence of a threshold exceedance when compared to the baseline data. The optimization problem is also constrained to ensure the same proportion of threshold exceedance measurements are made at each grid cell across the domain as determined with comparison to the baseline dataset. This constraint is included to maintain the spatial component of collection.

Our technique produced an ensemble of solutions varying with the number of monitors eliminated from the network. This research provides stake holders with information for: (1) The ideal monitor to remove from the network, (2) Which monitors are necessary and must be replaced if they fail, and (3) The identification of redundant monitors that can be relocated.

KEYWORDS: Air pollution, monitoring networks, optimized reduction methodology
1. INTRODUCTION
The Hamilton Air Monitoring Network (HAMN) consists of 14 monitors, of which 2 are operated by the Ontario Ministry of the Environment. Operating the 12 other devices comes at a large expense, which is currently funded by industrial partners located in Hamilton. However the commitment of partners is not insured and partners have left in the past, requiring the remaining partners to fill the funding gap (Annual Air Quality Report 2006 - HAMN). This air quality monitoring network (AQMN) is responsible for the monitoring of industrial emissions, defined by a diverse set of objectives. These objectives include: (1) Monitoring air emissions from industrial facilities to aid in the prevention of possible adverse effects; (2) Compliance with air quality regulations; (3) Providing accurate data for the use in air quality models; (4) Estimates of the local air quality emissions in adjacent communities and regionally, due to industrial outputs; (5) Provide data for an integrated air quality management control system; (6) Determine air quality improvement due to emission abatement strategies; and (7) Provide data to assess local population and ecosystem exposure to air pollutants. These objectives can be summarized as a single meta-objective, which is to identify when air quality is above prescribed levels (exceedance). At current, a time-series of data starting in 2003 to present is available. According to the most recent report released from HAMN the majority of their funds are spent to support the annual operating costs of $473,000 (Annual Air Quality Report 2006 - HAMN), with 10% of the annual operating budget committed to purchasing new equipment or conducting special air monitoring surveys. Thus, this monitoring network is very dependent on sustained funding from industrial partners. With Hamilton's declining traditional industrial sector (Adams et al., 2012); the potential for partners to revoke funding is present. To deal with possible monitor reduction due to reduced funding or operational failure, a methodology should be available to easily identify each monitor’s value to the network.

We present a methodology to evaluate the scenario of reduced monitor operation. A combinatorial optimization problem is formulated to identify the monitor that could be removed from the network by minimizing information loss about the air quality across the study area. Our approach compares interpolated air pollution surfaces of historic hourly concentration data obtained from the full set of air pollution monitors, with air pollution fields interpolated with data from reduced monitoring configurations. The focus is to identify the optimal reduced configuration that can identify air pollution concentrations that exceed the imposed air quality guidelines across the spatial extent of the study area.

2. METHODS

2.1 Study Area and Pollution Monitoring Data
The study area is located in Hamilton, Ontario, Canada. This area primarily consists of industrial land use, but is surrounded with residential property. It is located in the lower city; Hamilton consists of an upper and lower city separated by a 90 m escarpment. In the lower city over 25% of residents report their air quality perception as fair/poor (Simone et al., 2012). Hamilton is currently undergoing development that is resulting with spatial change in air quality (Adams et al., 2012). In figure 1 we present the study area, monitoring locations, a regular grid of locations at which the air pollution concentrations are assessed, and the industrial and non-industrial land use areas.
Within the study area eight stationary monitors operate measuring particulate matter 10 µm or less in aerodynamic diameter (PM$_{10}$) with tapered element oscillating microbalances (TEOM). The data we are using is a time-series from these eight monitors beginning on January 1st, 2011 and finishing on December 31st, 2011. These data are collected as one-hour average concentrations. The Ontario Ministry of the Environment has set PM$_{10}$ air quality criteria at 50 µg m$^{-3}$ [24-hour average] (Standards Development Branch Ontario Ministry of the Environment, 2012); this concentration standard will be referred to as PC. A 24-hour average was applied to the air pollution time-series, which were the values used in the optimization. To obtain the 24-hour averages, we included the previous 23 hourly measurements made in 2010 on December 31st to be able to obtain these values for the first 23 hours at the beginning of our time-series. The moving average $R'$ was determined with the following formula:

$$R' = \frac{c_h + c_{h-1} + \cdots + c_{h-23}}{24}$$

(1)

Where $c$ is the concentration obtained at the monitor, $h$ is the hour at which the moving average is being determined for.

2.2 Combinatorial Problem

A combinatorial problem was formulated to select a sub-set of monitors from the full set to maximize the retention of correctly identified air quality exceedance data obtained across the domain, when compared to the information obtained by the full monitor set. These comparisons are made at locations along a regular grid with 200 m spacing. The problem is constrained to ensure the same proportion of exceedance measurements annually are obtained at each grid cell across the domain as was with the full monitor set. This constraint is included to maintain the spatial representation of elevated concentrations. We have presented the solution for the reduction of one monitor from the network. We now present the formulation of the problem.

Let D be a regular grid of $n$ points located over the domain; $D = \{d_1, \ldots, d_n\}$. The spatial extent of D was obtained from the convex hull obtained with a 500 m buffer surrounding all monitoring stations. D locations are a 200 m regular grid of points; grid points located
The full set of monitors is \( G \) of \( q \) length, \( G = \{ g_1, \ldots, g_q \} \). For each of the monitors in \( G \), \( g_q \) has a time-series \( T \) of values of length \( m \) associated to it \( g_q = \{ g_{q1}, \ldots, g_{qm} \} \). \( T = \{ t_1, \ldots, t_m \} \). All \( T \) are identical and \( t_1 \) is the first hour of January 1st 2011, continuing to \( t_m \) which is the last hour of December 31st 2011. \( T \) is composed of an hourly data set, using the 24 hour moving average determined with formula 1.

For each hour \( t \) in \( T \), all values in \( G \) for that \( t \) are used to interpolate concentrations by ordinary kriging to each location in \( D \). These concentrations were considered the ideal air pollution concentrations for the study area. For each \( d_n \) in \( D \) this produced a time-series \( D = \{ d_{n1}, \ldots, d_{nm} \} \). For each \( t \) in \( T \) the error across the kriging surface is \( E, E = \{ e_1, \ldots, e_m \} \). Error was determined by leave-one out cross validation and averaged across all monitors for each \( t \).

Ordinary kriging was used for interpolation and conducted with the “Automap” package (Hiemstra et al., 2009), for R: A Language and Environment for Statistical Computing (R Core Team, 2012). Automap is an automated kriging process that was developed for interpolating air pollution fields.

A binary indicator approach is applied to identify when air pollution concentrations interpolated for each \( d_n \) are above the PC. These binary indicators are \( Z \), which is of \( n \times m \) length, \( Z = \{ z_{11}, \ldots, z_{nm} \} \).

\[
Z_{nt} = \begin{cases} 
1 & \text{if } d_{nt} \geq PC \\
0 & \text{otherwise} 
\end{cases} \quad (2)
\]

The total number of grid point concentration exceedances in \( Z \) for all \( t \) in \( T \) is \( \theta \):

\[
\theta = \sum_{n \in D} \sum_{t \in T} Z_{nt} \quad (3)
\]

The proportion of concentrations above PC for each \( z_n \) in \( Z \) of \( \theta \) is:

\[
Y_{dn} = \frac{\sum_{t \in T} Z_{t} \in Z_{dt}}{\theta} \quad (4)
\]

\( Y_{dn} \) is a value for each grid point location which is the proportion of exceedances measured throughout the year at that grid point from the entire sum of all grid points’ exceedance counts. The above \( Z \) and \( Y_{dn} \) represent the baseline information that each potential monitoring subset will be compared to, this will be referred to as \( \Omega \). The formulation of the comparison is now presented.

Potential subsets \( G' \) of \( G \) monitors will be of \( q \) length. For each \( G' \) we interpolate the concentrations at the \( D \) grid locations, which will be \( DS \). Interpolation is the same kriging procedure. \( DS \) is of \( q \times n \times m \) length, \( DS = \{ d_{11}, \ldots, d_{qnm} \} \) and are the values obtained when kriging fields based on the \( G' \) monitors. For each \( d_{snt} \) let \( X \) be indicator variable of \( n \times m \) length for when \( d_{snt} \) is above or equal to PC. \( X = \{ x_{11}, \ldots, x_{nm} \} \).

\[
X_{snt} = \begin{cases} 
1 & \text{if } d_{snt} \geq PC \\
0 & \text{otherwise} 
\end{cases} \quad (5)
\]

The total number of grid cell measurements for all \( d_{sn} \) in \( DS \) that exceed the PC for all \( t_n \) in \( T \) is \( \Phi \). \( \Phi \) is the total number of exceedances above the air quality guideline identified for all point locations for all hours of study.

\[
\Phi = \sum_{n \in D} \sum_{t \in T} X_{nt} \quad (6)
\]

The proportion of concentrations above PC in each \( d_{sn} \) in \( DS \) of \( \Phi \) is:
\[ \lambda_{s\text{dn}} = \frac{\sum_{t \in \Omega} X_t \in X_{dt}}{\Phi} \]  

(7)

For each G' set of results we have an \( X_{s\text{nt}} \) and \( \lambda_{s\text{dn}} \) which will be compared to Z and \( Y_{dn} \) respectively from \( \Omega \). The comparison will use another indicator variable that can take on one of three values, 1, 0 or -1. Correct values of \( d'_{s\text{nt}} \) compared to \( d_{nt} \) when concentrations are above PC is given a value of 1; -1 for incorrect values of \( d'_{s\text{nt}} \) for \( d_{nt} \) when concentrations are above PC, or when \( d'_{s\text{nt}} \) values are greater than PC but \( d_{nt} \) values were not; and 0 for incorrect or correct prediction when \( d_{nt} \) is below PC, as these values are of lesser importance because they are not in violation of air quality guidelines.

\[ Y = \{y_1, ..., y_n\} \]. There is a range of acceptable values for \( d'_{s\text{nt}} \) determined by the \( e_t \) for kriging error values in \( \epsilon \).

\[
y_{nt} = \begin{cases} 
1 & \text{if } d'_{s\text{nt}} = d_{nt} \pm e_t, \text{ when } d_{nt} \geq PC \\
-1 & \text{if } d'_{s\text{nt}} \neq d_{nt} \pm e_t, \text{ when } d_{nt} \geq PC \\
-1 & \text{if } d'_{s\text{nt}} > PC \pm e_t, \text{ when } d_{nt} < PC \\
0 & \text{else} 
\end{cases} \]  

(8)

To determine which monitor can be removed from the network with the least overall effect to the number of air pollution exceedances, we maximize the value of \( y_{nt} \) with the formula specified below:

\[
\text{Maximize } (Y) = \sum_{d=1}^{n} \sum_{t=1}^{m} y_{nt} 
\]  

S.T.

\[
Y_{dn} = \lambda_{d\text{dn}} \pm \alpha \forall d_n \in D 
\]  

The constraint is included to ensure the spatial distribution of exceedances is retained by ensuring that the proportion of total exceedances estimated for all of the grid point location is within \( \pm \alpha \) when the values interpolated for the monitor sub-sets are compared to the entire monitor set. \( \alpha \) is a specified amount for how much the proportion of elevated concentrations occurrences can vary. In this study \( \alpha \) was set to a value of 0.005, which would require no grid point to vary greater than 0.5 in proportional exceedances measured.

2.3 Pseudo Code for Optimization

The optimization program was written and executed in R: A language and environment for statistical computing. Pseudo code is presented below.

### For full monitoring set

For each time in the time-series:

- Interpolate an air pollution surface by ordinary kriging with the full monitor set
- Assign air pollution concentrations to each grid point
- Save the kriging error value
- For each cell in the grid:
  - Assign a 1 or 0 value to the indicator variable for concentrations above PC
Sum all the PC indicator variables and save as overall total
For each cell in the grid:
- Sum the number of indicator variable for exceedances and divide by the overall total.

### For Subsets of monitors

For each potential subset of monitors:

- For each time in the time-series:
Interpolate a concentration surface with kriging of the monitor sub-set and assign values to each grid cell

For each cell in the grid:
Assign a 1 or 0 value to the indicator variable for concentrations above PC

Sum all the PC indicator variables
For each cell in the grid:
Sum the individual PC indicator variables divide by the total sum.
For each cell in the grid:
Determine if the value compared to the full monitor set is correctly identified, a false-positive, or a false negatives

Validate solution if proportion of PC concentrations is suitable in each cell

### To obtain the solution
Select solution with Maximum $Y$-score.

3. RESULTS

The maximum 24-hour averaged concentration across all air pollution monitors was 122 µg m$^{-3}$, which is 244% of the air quality guideline. The number of hours with at least one of the eight monitors above PC is 162 hours of the 8,760 hours of the year (1.85%). The proportion of monitors recording exceedances by hour is presented in Figure 2. The interpolated values at the point locations were estimated to exceed the air pollution guideline concentration of 50 µg m$^{-3}$ 56,184 times of the 3,626,640 total interpolated values (414 monitors * 8760 hours).

![Figure 2. The number of monitors’ exceedances by hour.](image)

With each of the potential monitor subset, each of the 414 grid points were evaluated as correctly identified, false-positive, or false-negative in regards to air quality exceedance above the air quality guideline. Given the exceedance data from the full monitoring set, the highest potential Y-value was 56,184, and if a monitoring configuration attained this value with a monitor removed it would suggest that the removed monitor provides no new information. The value of 56,184 was the number of exceedances measured by the full monitoring network. The actual Y-scores obtained for the monitor subsets by individually removing a monitor ranged from 24,605 to 54,008.

The maximum $\alpha$ value, which was the proportional change in the number of exceedances identified at a monitor over the year compared to the total, showed little variability between solutions, with a range from 0.0010 to 0.0012. All solutions were produced with a maximum absolute difference for $\alpha$ well within the allowable limit specified at 0.005. We present the Y-score for each potential solution, and also the absolute max proportion change ($\alpha$) in Table 1. We plotted Y against the maximum $\alpha$ (not presented), no relationship existed.

CEST2013_0599
4. DISCUSSION

Air pollution optimal monitoring network design research has three primary approaches. The first is the geostatistical approach (Baalousha 2010; Haas 1992; Nunes et al., 2006; Trujillo-Ventura and Ellis, 1991), which attempts to locate monitors by minimizing kriging error. A second approach utilizes the concept of entropy to locate monitoring devices (Ainslie et al., 2009; Caselton et al., 1992). Monitors are located in a way that they capture the most information from the field. The third approach is to define an objective function to minimize or maximize. An example includes maximizing the number of violations (Modak and Lohani, 1985). We present an optimization approach that utilizes the interpolation error into its results, but is designed to optimize an objective function that maximizes the retention of records about the number of air quality guideline exceedances that occur.

Table 1. Results from Reduced Monitoring Configurations, Y value and Maximum α.

<table>
<thead>
<tr>
<th>Removed Monitor ID</th>
<th>Y</th>
<th>Maximum α</th>
</tr>
</thead>
<tbody>
<tr>
<td>29153</td>
<td>54,008</td>
<td>0.0011</td>
</tr>
<tr>
<td>29170</td>
<td>53,596</td>
<td>0.0012</td>
</tr>
<tr>
<td>29113</td>
<td>47,950</td>
<td>0.0011</td>
</tr>
<tr>
<td>29102</td>
<td>45,999</td>
<td>0.0011</td>
</tr>
<tr>
<td>29567</td>
<td>45,873</td>
<td>0.0012</td>
</tr>
<tr>
<td>29565</td>
<td>38,677</td>
<td>0.0010</td>
</tr>
<tr>
<td>29168</td>
<td>37,038</td>
<td>0.0011</td>
</tr>
<tr>
<td>29154</td>
<td>24,605</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

Spatial distribution of the air quality exceedances are maintained in our optimization program by constraining results to require similar proportions (+/- α) of identified exceedances for each grid point across the study area. In our results this did not seem particularly important as there was very little variability, the proportions only varied between all subset results by 0.0002. Our study area that we have applied this technique to is quite small and we believe that applying this approach to a larger study area with multiple sources spread apart would increase the importance of this constraint.

We found that two of the monitors would be able to be independently removed with very little reduction in the number of air pollution exceedances that would be correctly identified. Monitors 29153 and 29170 can both be removed with a remaining accuracy of 96 and 95% respectively. This slight loss of information should not drastically affect the information obtained about the network.

This approach can provide monitoring network advisors information about the importance of each particular monitor regarding air quality guideline exceedances across the study area and allow them to make informed decisions as to whether a monitor failure must be corrected immediately. An example would be the failure of monitor 29154 in the Hamilton network, with this monitor failing only 44% of the exceedances would have been identified. This monitor has a high value within the network, which is likely due to its location as the most northerly monitor. Being located to the north, it is the only monitor well aligned to capture air pollution that is emitted from the industrial sector when the wind is blowing from the south; particularly the south-east. Winds coming from the south-west are the primary wind direction in Hamilton, ON (Arain et al., 2009). If monitoring cannot be appropriately conducted during southwest winds, the polluters in the area may go unnoticed affecting a few hundred households located near monitors 29102 and 29154. These households already experience elevated air pollution concentrations as they are located directly beside a major freeway in the area, which is a gateway between...
the Greater Toronto Area and New York State in the United States of America (Adams et al., 2012).

5. CONCLUSIONS

We have presented a monitor evaluation technique to determine which monitors in a network can be removed with little impact to the quality of data provided on air quality guideline exceedances. A solution to the optimization problem has been presented for the removal of only one monitor, but the framework easily allows for the extension to simulate the removal of any combination of monitors. However, because of the interpolation processes, at minimum 3 monitors must be maintained and these monitors must be spatially distributed within the study area. We penalized false-positives and false-negatives equally, our program is easily adaptable to modify the penalization costs for false-positives or false-negatives, which would be modified in formula 8. The solution and process is dependent on the air pollution field having stationary processes. If the air pollution field changes dramatically year to year this process should be conducted for multiple years and then further interpretation of the time-series of the Y-scores can be made.

A major benefit from this approach is its simplicity to be adapted for any monitoring network. The only changes would be input data files containing the station locations, a grid of points to evaluate, the concentration standard, and the time-series of pollution data for each monitor. As well, it is a very fast process to run; currently it runs as a serial program and only takes a couple of hours to run through the entire process. The intention is to run interpolation procedure in parallel to reduce the overall run time.

REFERENCES


CEST2013_0599