Optimal Air Quality Monitoring Network for Green Cities

Mohammed Mujtaba Shareef¹, Tahir Husain¹, and Badr Alharbi²
¹Faculty of Engineering and Applied Science, Memorial University, St. John’s, NL, A1B 3X5, Canada
²National Center for Environmental Technology, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia

Abstract. Continuous efforts are being made by cities around the world to identify methods to make their environment healthier and more livable through the improvement of water, land and air. Air pollution is of high importance since it often causes severe and irreversible damage to human health and environment and remediation is extremely difficult. The potential sources of air pollution, such as large industries, automobiles and power plants, are essential part of cities. Transforming these cities into “green cities” necessitates control of these emissions. Continuous monitoring of air pollution with a well-designed air quality monitoring network (AQMN) is usually the first step in addressing and tackling the emissions. Environmental protection agencies are looking for an optimal designed AQMN with an obvious focus on minimizing costs, along with other objectives. This article proposes a simple method of optimizing the AQMN using geographical information system (GIS), interpolation techniques and historical data. Existing air quality stations are systematically eliminated, and the missing data is filled in with data generated from the most appropriate interpolation technique. The interpolated data is then compared with the observed data. Pre-defined performance measures were used to check the accuracy of the interpolated data. An algorithm was developed in a GIS environment and the process was simulated for several sets of measurements. This methodology proved to be useful to decision makers to find optimal numbers of stations that are needed without compromising the coverage of the concentrations across the city.

Keywords. air quality monitoring, GIS, green cities, interpolation, Kriging, IDW

1. Introduction

It is well known that air pollution causes adverse effects on human health and negatively impact the environment. Due to rapid urbanization and industrialization, air pollution is of high significance, particularly in large cities, when it comes to keeping the city green. Continuous monitoring of air pollution with a well-designed air quality monitoring network (AQMN) is the first step in addressing this issue. Obtaining continuously monitored data to ensure safe levels of air quality is one of the primary objectives of AQMN in addition to evaluating exposure hazards and implementing effective control strategies. Environmental protection agencies are looking for an optimal design of AQMN meeting these objectives with an obvious focus on minimizing cost.

The methodology used to design a new AQMN or evaluate an existing AQMN attracted the attention of several researchers. Maximum sensitivity of the collected data (Koda and Seinfeld, 1978; Caselton and Husain, 1980) and maximum coverage factors such as intensity of emissions, source distance and meteorology (Houghland et al., 1980) were some of the first techniques used to design an AQMN. Statistical measurements of information content (Picket and Whiting, 1981) and Fisher’s information measure (Husain and Khan, 1983) was used to determine the optimum number and location of monitors in a network. The design of an AQMN in the greater London area was conducted by Handscombe and Elson
(1982) based on the concept of a spatial correlation analysis. The same concept combined with potential of violation was used by Arbeloa et al., (1993) to design an optimal air quality monitoring network. Noll and Mitsutome (1983) developed a method to design AQMN based on expected ambient pollutant dosage. This method ranks potential locations by calculating the ratio of a station’s expected dosage over the study area’s total dosage. Another methodology developed by Nakamori and Sawaragi (1984) determines the representative areas of monitor stations in urban areas. A different perspective, based on the use of Shannon information, was initiated with the results of Caselton and Zidek (1984), applied in a univariate setup by Sampson and Guttorp (1992) and Guttorp et al., (1993) and later extended to a multivariate context by Perez-Abreu and Rodriguez (1996). The concept of sphere of influence and figure of merit was applied by McElroy et al., (1986) to calculate the minimum number of air quality monitoring sites. A simple methodology for siting ambient air monitors at the neighborhood scale was developed by Richard et al., (2002) and applied as a case study in Hudson County.

A linear programming approach was used by many researchers to site optimum AQMN. A multi-attribute utility function method was used for siting the air quality network by Kainuma et al., (1990). Trujillo-Ventura and Ellis (1991) applied multiple objectives, including spatial coverage, violation detection, data validity and a weighting method, to determine the most suitable network for Tarragona, Spain. Chen et al., (2006) developed a multiple objective optimization model for air quality monitoring networks in Taoyuan County, Taiwan. A multiple objective optimization model and a procedure for sustainable air quality monitoring networks planning were developed. A holistic approach was adapted for optimal design of air quality monitoring network expansion in an urban area by Mofarrah and Husain (2010). In this approach a multiple-criterion method with the spatial correlation technique was implemented to design an expanded air quality monitoring network of Riyadh city in Saudi Arabia. Tseng and Ni-Bin (2001) proposed a Genetic Algorithm (GA) based compromise programming technique for assessing the relocation of strategy of an urban air quality monitoring network with respect to the multi-objective and multi-pollutant design criteria. Silva and Quiroz (2003) used a Shannon information index to optimize atmospheric monitoring networks by excluding the least informative stations with respect to different air pollutants. An optimal design of AQMN was conducted around a refinery using a mathematical model based on the multiple cell approach with simultaneous multiple pollutants by Elkamel et al., (2008). Lu et al., (2011) used principal component analysis (PCA) and cluster analysis to optimize the air quality monitoring network in Hong Kong. PCA and fuzzy c-means clustering was applied for assessment of air quality monitoring in Turkey by Dogruparmak et al., (2014). The authors showed that a number of monitoring stations can be decreased using the methodology. Ferradas et al. (2011) developed a methodology based on a self-organizing map (SOM) artificial neural networks for integrating data about multiple measured pollutants to group monitoring stations according to their similar air quality. The proposed method considered the subsequent geographical mapping of the clusters of stations observed with the SOM, which made it possible to detect geographically different areas that share similar air pollution problems.

The strides that the field Geographical Information System (GIS) and its components (such as interpolation methods) are making as an application in almost every field are incredible. GIS and spatial interpolation techniques were also used in AQMN. Bayraktar et al., (2005) used a Kriging-based approach to locate sampling sites for assessing air quality. Long years of the smoke data measurements were used in the determination of the region for the representation by drawing contours through the Kriging method. Then the sampling site in this region was selected based on the EPA criteria. External drift Kriging of NOx concentrations with dispersion model output was used by Kassteele et al., (2009) to reduce the uncertainties in parameter estimation due to the reduction of air quality networks. GIS ancillary variables were used to predict volatile organic compounds and nitrogen dioxide levels at unmonitored locations (Smith et al., 2006).
The predictive equations were developed by regressing the passive monitor measurements at the 22 monitored schools on land-use variables derived from GIS. A universal Kriging interpolation method was found to perform better than other methods when a comparison was made between various interpolation methods by developing EU-wide high resolution air pollution maps (Beelan et al., 2009). The best method to model concentrations was selected on the basis of predefined performance measures, correlation coefficient and RMSE. A generalized monotonic regression based on B-splines with an application to air pollution data was proposed by Leitenstorfer and Tutz (2007) and the spline method was used to estimate the varying health risks from air pollution across Scotland by Duncan (2011).

The methods summarized above are very useful, well established and have been implemented widely. However, it appears a simple GIS based methodology would further reduce the complexities of AQMN design. The basic advantage of using GIS is that it organizes geographic data in such a way that the decision making process becomes easy. In addition to this, it provides several advanced functionalities to manage statistical and spatial data, interpolate the data to create smooth surface, extract data from the interpolated surface and create algorithms to automate the process. Furthermore, it creates results that can be visualized in interactive maps, which will further simplify the decision making process. Taking the cue on these advantages, this paper proposes a simple and innovative process to optimize AQMN by using GIS, interpolation methods and historical data. The existing stations are systematically eliminated by creating several interpolated maps and comparing it with the observed values. The number of stations that can be eliminated is governed by the pre-defined performance measures criteria. Recently an increasing trend of air pollution has been observed in Riyadh city of Saudi Arabia and there is an emphasis on frequent air pollution measurements (Alharbi et al., 2015), hence Riyadh is used as a case study to test the proposed methodology.

2. Methodology

2.1 Interpolation Methods

Interpolation predicts values for cells in a raster using a limited number of sample data points, which helps in predicting unknown values for any geographic station. Five interpolation methods were selected to estimate the concentrations of air pollution at the unknown stations. The selected methods were (1) Inverse Distance Weighted (IDW), (2) Spline (SPL), (3) Ordinary Kriging (OK), (4) Universal Kriging (UK) and (5) Natural Neighbor (NN). These methods have been widely used in estimating air pollution concentrations.

The IDW uses a method of interpolation that estimates cell values by averaging the values of sample data points in the neighborhood of each processing cell (ESRI Redland). The closer a point is to the center of the cell being estimated, the more influence or weight it has in the averaging process. This has been used in several instances to interpolate air pollutant concentrations (Wong et al., 2004; Hoek et al., 2002; Jerrett et al., 2013).

The SPL uses an interpolation method that estimates values using a mathematical function that minimizes overall surface curvature, resulting in a smooth surface that passes exactly through the input points. This method was found superior in varying health risk from air pollution (Duncan, 2012) and was also recommended by Leitenstorfer and Tutz (2006).

The Kriging is an advanced geo-statistical procedure that generates an estimated surface from a scattered set of points with z-values. While Kriging is a weighted combination of monitor values, this method also uses spatial auto correlation among data to determine the weights. Generally Kriging has two different forms—ordinary and universal Kriging (Burrough, 1998). In the ordinary Kriging the mean value is assumed constant and determined during interpolation, and universal Kriging assumes that data follows a known trend. Ordinary and universal Kriging have previously been used with success to model ozone (Liu and Rossini, 1996) and particles (Cressie, 2006) at the local scale and to model broad scale variations in background air pollution (Lefohn et
Kriging has also been implemented in several other studies (Beelan et al., 2009; Zou et al., 2011; Shad et al., 2009).

Natural Neighbor interpolation finds the closest subset of input samples to a query point and applies weights to them based on proportionate areas to interpolate a value. It is also known as Sibson or “area-stealing” interpolation.

2.2 Performance Measures

A statistical error is the amount by which an observation differs from its expected value. The statistical indices selected to measure performance are Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Nash-Sutcliffe equation (NSE) (Nash and Sutcliffe, 1970) and Accuracy Factor (ACFT) (Ross, 1996) given by equations 1 to 4, respectively. The Pearson correlation coefficient \( r \) was also measured to check the strength and direction of the linear relationship between the observed and interpolated values.

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \text{Intr}_i - \text{Obs}_i \right)^2} \quad [1]
\]

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left( \left| \frac{\text{Intr}_i - \text{Obs}_i}{\text{Intr}_i} \right| \right) \quad [2]
\]

\[
\text{NSE} = 1 - \frac{\sum_{i=1}^{n} \left( \text{Intr}_i - \text{Obs}_i \right)^2}{\sum_{i=1}^{n} \left( \text{Intr}_i - \text{Obs}_i \right)^2} \quad [3]
\]

\[
\text{ACFT} = 10 \left( e^\left( \frac{\left| \log \left( \frac{\text{Intr}_i}{\text{Obs}_i} \right) \right|}{n} \right) \right) \quad [4]
\]

Where \( \text{Intr}_i = \) Interpolated value; \( \text{Obs}_i = \) Observed value; \( n = \) number of observations

RMSE is the frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. It basically represents the sample standard deviation of the differences between predicted and observed values. RMSE gives important information in predicting the magnitude of a pollutant concentration, a measure close to zero represents good predictions. The absolute mean percentage error denoted by MAPE is calculated by dividing the sum of percentage error by the number of observations, and a value equal or close to zero is considered ideal. The correlation coefficient is a measure of linear dependence between two variables, and it was chosen to get an indication of the correspondence of timing and evolution between observed and interpolated concentration values. The coefficient of efficiency (NSE) indicates the normalized fit of the model, the value ranges from \(-\infty\) to 1. It compares the mean square error generated by a particular model simulation to the variance of the output sequence; a value of 1 indicates a perfect fit (Nash and Sutcliffe, 1970). Ross, 1996 proposed a simple multiplicative factor called accuracy factor (ACFT) that represents the spread of the results of the modelled data. A value of one indicates that there is perfect agreement between all the predicted and modelled values and values larger than one indicate the less accurate average estimate.

Several studies have used error statistics to compare observed and predicted meteorology and air quality data. RMSE was used by Shad et al. (2009) to compare the observed data with the predicted air pollution using fuzzy genetic linear Kriging. Monteiro et al., (2013) used it in bias correction techniques to improve air quality ensemble predictions, and Son et al., (2010) used it in a study of individual exposure to air pollution and lung function in Korea. EU-wide maps were prepared based on the predictor variables and the modelled air pollution concentrations were selected on the predefined performance measures \( r^2 \) and RMSE. A value of RMSE <10 and \( r^2 >0.5 \) was considered a good performance measure (Beelan et al., 2009). Singh et al., (2012) used RMSE, ACFT and NSE as performance measures of different linear and nonlinear modelling approaches for predicting urban air quality. Solar radiation and the air pollution index were estimated based on linear, exponential and logarithmic models, and similar indices were used to check the performance index in China (Zhao et al., 2013). In the current study, RMSE, MAPE and \( r^2 \) were primarily used as performance measures. RMSE <8 MAPE <25% and \( r^2 >0.5 \) was considered a good measure.

2.3 Station Elimination Process

The main objective of this optimization process is to eliminate as many stations as possible and fill in the missing information through the interpolated values. The steps involved in this process...
are illustrated in Figure 1 and elaborated as follows.

2.3.1 Step 1: Selection of Stations
As a first step, a single station (P) or set of stations (P₁, P₂...) are selected to be eliminated from the vector dataset used for creating the raster. A particular station or set of stations can be chosen or a set of all possible stations tested in a loop.

2.3.2 Step 2: Storing the Observed Values
The observed concentrations at the selected station (P()) or stations (P₁(), P₂() ...) are stored as arrays. These values will later be compared with the interpolated values.

![Flowchart](image)

**Figure 1:** Station elimination process
2.3.3 Step 3: Creating Vector Layer
A vector layer is a coordinate-based data model that represents geographic features, such as points. Each point is the station represented by the geographical coordinates. The layer has a column of “z” values used for interpolation. Several columns of “z” values, such as the concentration of the pollutants, are added to the vector for simulation. In this step, the vector layer is created without the selected stations.

2.3.4 Step 4: Creating Raster from the Vector
Raster is defined as a spatial data model that defines space as an array of equally sized cells arranged in rows and columns and composed of single or multiple bands. Each cell contains an attribute value and location coordinates. Rasters are created using the vector data and by applying appropriate interpolation techniques. In this step, five rasters are created using IDW, Spline, Ordinary Kriging, Universal Kriging and Natural neighbor methods.

2.3.5 Step 5: Extracting the Interpolated Value
This step calculates the predicted concentrations. The values at the eliminated stations are extracted from the created rasters. The values for each interpolation method (IDW, SPL, OK, UK, and NN) are stored in separate arrays (IDW1(), IDW2(); SPL1(); SPL2();...; OK1(); OK2();...; UK1(); UK2();... and NN1(); NN2();...). This process of interpolation and value extraction is repeated for all the observed datasets.

2.3.6 Step 6: Performance Measure
The process of interpolation and value extraction generates arrays of the interpolated values. In this step, the performance measures are applied to observed and interpolated values. RMSE, Bias, and correlation coefficient are calculated for the selected stations and the five interpolation methods. The interpolation method that generates the minimum performance measure is then chosen and the others are discarded. These
measures are compared with the pre-defined threshold limits, and if the measures are within the limit, they are stored in the possible station combinations array C().

2.3.7 Step 7: Repeat for another Station Combination

The process is repeated for another station combination until all the combinations are exhausted.

2.3.8 Step 8: Finding the Best Possible Station Combination

The best possible station combinations can be chosen from the list of possible station combinations array C(). The decision maker can then choose from the list of possible station combinations, which can be eliminated from the AQMN.

2.4 Study Location and Field Measurements

The proposed methodology is applied to the city of Riyadh, Saudi Arabia. The city is divided into sixteen cells that are identical in area, and each cell is 12 km x 12 km. The measurements were carried out intermittently from September 2011 to September 2012. Most of the measurements have been conducted approximately in the center of each cell (Figure 2) with two equipped mobile air quality monitoring stations capable of monitoring meteorological variables as well as CO, O₃, NOₓ, CH₄, OC, EC and PM₂.₅. However, the methodology suggested in this study is implemented for the following four criteria pollutants: SO₂, NOₓ, O₃ and CO. The type of sensors used with their respective method of monitoring for the pollutants utilized in this study are NO, NO₂ and NOₓ Chemiluminescence; CO-Dual Beam NDIR; O₃-UV Photometer and SO₂- UV Fluorescence.

As the measurements are staggered, in order to get a continuous dataset, 24 datasets were prepared from the available measurements by averaging the hourly measurements for the entire study period for all the 16 stations. These 24 datasets were used for the simulation to create the raster with different interpolation techniques and compared with the observed values. ESRI’s ArcGIS (ESRI, Redlands) exposes several functions to run the interpolations and extract the necessary data. These functions were customized to run the simulation process through Microsoft Visual Basic for application (VBA) environment in ArcGIS.

3. Results and Discussion

The pollutant concentration data was collected from 16 stations as shown in Figure 2. It is assumed that a minimum of eight stations are needed to produce reliable interpolated maps. Therefore, up to eight possible elimination scenarios were simulated. Table 1 outlines the simulation details such as the number of stations to be eliminated, possible number of combinations, number of simulations and the approximate time of simulation (on a high end PC). As seen in the table, the station combinations and the computing power required increase with the increase in number of stations. The methodology proposed in this study is tested on the four pollutants O₃, NOₓ and SO₂ and CO, and the results of the simulations are discussed as follows. The interpolation was performed using the IDW, Spline, OK, UK and NN methods. The IDW and UK outperformed other methods, particularly in terms of producing lower RMSE, MAPE values and a higher value of r².

Table 1

<table>
<thead>
<tr>
<th>No. of stations to be eliminated</th>
<th>Possible combinations</th>
<th>No. of simulations</th>
<th>Approximate time of simulation (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>1,920</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>120</td>
<td>14,400</td>
<td>2.25</td>
</tr>
<tr>
<td>3</td>
<td>560</td>
<td>67,200</td>
<td>10.5</td>
</tr>
<tr>
<td>4</td>
<td>1,820</td>
<td>218,400</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>4,368</td>
<td>524,160</td>
<td>81</td>
</tr>
<tr>
<td>6</td>
<td>8,008</td>
<td>960,960</td>
<td>150</td>
</tr>
<tr>
<td>7</td>
<td>11,440</td>
<td>1,372,800</td>
<td>214</td>
</tr>
<tr>
<td>8</td>
<td>12,870</td>
<td>1,544,400</td>
<td>241</td>
</tr>
</tbody>
</table>
Table 2
Performance measures for O₃ simulations

<table>
<thead>
<tr>
<th>No. of station(s) to be eliminated</th>
<th>Optimal station(s) combination that can be eliminated</th>
<th>Interpol. method</th>
<th>RMSE</th>
<th>r²</th>
<th>MAPE</th>
<th>NSE</th>
<th>ACFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11, 15</td>
<td>IDW</td>
<td>3.224</td>
<td>0.953</td>
<td>4.041</td>
<td>0.944</td>
<td>0.960</td>
</tr>
<tr>
<td>2</td>
<td>11, 15</td>
<td>IDW</td>
<td>3.811</td>
<td>0.925</td>
<td>5.153</td>
<td>0.919</td>
<td>0.973</td>
</tr>
<tr>
<td>3</td>
<td>4, 11, 15</td>
<td>UK</td>
<td>3.964</td>
<td>0.913</td>
<td>7.991</td>
<td>0.901</td>
<td>1.003</td>
</tr>
<tr>
<td>4</td>
<td>4, 11, 12, 15</td>
<td>UK</td>
<td>4.588</td>
<td>0.911</td>
<td>15.699</td>
<td>0.884</td>
<td>1.099</td>
</tr>
<tr>
<td>5</td>
<td>4, 5, 11, 15, 16</td>
<td>IDW</td>
<td>5.495</td>
<td>0.862</td>
<td>22.805</td>
<td>0.844</td>
<td>1.095</td>
</tr>
<tr>
<td>6</td>
<td>4, 8, 11, 12, 13, 15</td>
<td>UK</td>
<td>6.113</td>
<td>0.831</td>
<td>25.322</td>
<td>0.821</td>
<td>1.015</td>
</tr>
<tr>
<td>7</td>
<td>4, 5, 8, 11, 12, 15, 16</td>
<td>IDW</td>
<td>6.488</td>
<td>0.832</td>
<td>36.469</td>
<td>0.809</td>
<td>1.081</td>
</tr>
<tr>
<td>8</td>
<td>4, 5, 6, 8, 11, 12, 15, 16</td>
<td>IDW</td>
<td>7.018</td>
<td>0.817</td>
<td>49.610</td>
<td>0.775</td>
<td>1.138</td>
</tr>
</tbody>
</table>

Figure 3: Variation of RMSE with respect to combination of stations that are eliminated

3.1 O₃
Table 2 presents the results of the simulation for O₃. The values of RMSE, r², MAPE, NSE and ACFT are shown along with the number of stations to be eliminated. The variations of RMSE, MAPE and r² with respect to stations are illustrated in Figure 3, Figure 4 and Figure 5, respectively. Station 11 was the best single station that could be eliminated, and stations 4, 5, 6, 8, 11, 12, 15 and 16 were the best eight stations. RMSE and MAPE increased from 3.224 to 7.018 and 4.041 to 49.61 from one station elimination to eight, respectively. Similarly, NSE and ACFT decreased from 0.944 to 0.775 and 0.960 to 1.138, respectively. In order to limit MAPE within 25%, the maximum number of stations that could be eliminated was six (4, 8, 11, 12, 13 and 15). This also generated a satisfactory RMSE (6.113) and r² (0.831).
Figure 4: Variation of MAPE with respect to combination of stations that are eliminated

Figure 5: Variation of $r^2$ with respect to combination of stations that are eliminated
Figure 6: Stations required to generate \( O_3 \) concentration maps (RMSE<7 and MAPE<25)

Table 3: Performance measures for NO\(_x\) simulations

<table>
<thead>
<tr>
<th>No. of station(s) to be eliminated</th>
<th>Optimal station(s) combination that can be eliminated</th>
<th>Interpol. method</th>
<th>RMSE</th>
<th>( r^2 )</th>
<th>MAPE</th>
<th>NSE</th>
<th>ACFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>IDW</td>
<td>5.999</td>
<td>0.865</td>
<td>3.792</td>
<td>0.848</td>
<td>1.020</td>
</tr>
<tr>
<td>2</td>
<td>3, 11</td>
<td>IDW</td>
<td>7.470</td>
<td>0.808</td>
<td>4.006</td>
<td>0.731</td>
<td>0.961</td>
</tr>
<tr>
<td>3</td>
<td>3, 5, 11</td>
<td>UK</td>
<td>9.084</td>
<td>0.619</td>
<td>6.591</td>
<td>0.599</td>
<td>0.984</td>
</tr>
<tr>
<td>4</td>
<td>3, 5, 7, 11</td>
<td>UK</td>
<td>10.766</td>
<td>0.439</td>
<td>12.184</td>
<td>0.430</td>
<td>1.025</td>
</tr>
<tr>
<td>5</td>
<td>3, 5, 7, 11, 14</td>
<td>IDW</td>
<td>11.521</td>
<td>0.461</td>
<td>15.621</td>
<td>0.459</td>
<td>1.022</td>
</tr>
<tr>
<td>6</td>
<td>3, 5, 7, 11, 14, 16</td>
<td>UK</td>
<td>12.119</td>
<td>0.438</td>
<td>19.839</td>
<td>0.438</td>
<td>1.019</td>
</tr>
<tr>
<td>7</td>
<td>3, 5, 6, 7, 11, 14, 16</td>
<td>IDW</td>
<td>14.834</td>
<td>0.395</td>
<td>23.148</td>
<td>0.341</td>
<td>0.966</td>
</tr>
<tr>
<td>8</td>
<td>1, 3, 5, 6, 7, 11, 14, 16</td>
<td>IDW</td>
<td>14.684</td>
<td>0.312</td>
<td>27.468</td>
<td>0.287</td>
<td>0.977</td>
</tr>
</tbody>
</table>

The accuracy factor (ACFT) was 1.015, which is considered to be very good performance. The remaining 10 stations that are required to produce satisfactory maps of ozone are shown in Figure 6. These stations could essentially be considered hotspots of \( O_3 \) emissions.

The primary sources of ground level \( O_3 \) are automobiles, cement and power plants, construction activities and biogenic or natural sources. Small industries such as paint shops, dry cleaners and bakeries are also known to contribute \( O_3 \). These 10 stations fall within these areas. The stations 10 and 7 are located in a dense residential area, and stations 2 and 3 are in an industrial area. The agricultural areas are found near station 9, and construction activities are reported near station 14. Stations 1 and 16 are the city outskirts where small scale industries are located.

3.2 NO\(_x\)

The simulation results for NO\(_x\) concentrations are shown in Table 3. For the one station elimination scenario, it was found that station 11 was the best one with a RMSE=5.999; MAPE=3.792
and $r^2=0.864$. These values increased to 14.684, 27.468 and 0.312, respectively, for eight station eliminations (Figures 3, 4 and 5). For MAPE scenario of $\leq 25$, 7, a maximum of seven stations could be eliminated (3, 5, 6, 7, 11, 14 and 16). However, RMSE and $r^2$ values were not within the predefined limits. Taking the MAPE as priority, a minimum of nine stations were needed to produce satisfactory NO$_x$ concentration maps as shown in Figure 7. In addition to the industrial emissions, the primary source of NO$_x$ is from automobile emissions, which are reasonably evenly distributed over the study area, and hence a fewer number of stations is required to produce the concentration maps. Stations 4 and 8 signify the presence of industries, and the other stations are spread over the densely populated areas in the city where the automobiles move predominantly. The ACFT=0.966 and NSE=0.341 supports the elimination of the seven stations.

![Figure 7: Stations required to generate NO$_x$ concentration maps (RMSE<15 and MAPE<25)](image)

<table>
<thead>
<tr>
<th>No. of station(s) to be eliminated</th>
<th>Optimal station(s) combination that can be eliminated</th>
<th>Interpol. method</th>
<th>RMSE</th>
<th>$r^2$</th>
<th>MAPE</th>
<th>NSE</th>
<th>ACFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>UK</td>
<td>2.155</td>
<td>0.323</td>
<td>2.155</td>
<td>-1.507</td>
<td>0.983</td>
</tr>
<tr>
<td>2</td>
<td>1, 4</td>
<td>IDW</td>
<td>3.991</td>
<td>0.446</td>
<td>5.925</td>
<td>0.421</td>
<td>1.029</td>
</tr>
<tr>
<td>3</td>
<td>1, 4, 13</td>
<td>UK</td>
<td>4.599</td>
<td>0.375</td>
<td>10.792</td>
<td>0.355</td>
<td>1.042</td>
</tr>
<tr>
<td>4</td>
<td>1, 4, 6, 13</td>
<td>IDW</td>
<td>4.737</td>
<td>0.268</td>
<td>15.562</td>
<td>0.167</td>
<td>1.063</td>
</tr>
<tr>
<td>5</td>
<td>1, 4, 6, 13, 15</td>
<td>IDW</td>
<td>5.124</td>
<td>0.582</td>
<td>25.187</td>
<td>0.428</td>
<td>1.140</td>
</tr>
<tr>
<td>6</td>
<td>1, 4, 6, 11, 13, 15</td>
<td>IDW</td>
<td>6.258</td>
<td>0.642</td>
<td>43.558</td>
<td>0.284</td>
<td>1.248</td>
</tr>
<tr>
<td>7</td>
<td>1, 4, 6, 7, 11, 13, 15</td>
<td>UK</td>
<td>7.579</td>
<td>0.469</td>
<td>60.337</td>
<td>0.038</td>
<td>1.278</td>
</tr>
<tr>
<td>8</td>
<td>1, 4, 6, 7, 9, 11, 13, 15</td>
<td>UK</td>
<td>9.962</td>
<td>0.301</td>
<td>69.931</td>
<td>0.100</td>
<td>1.279</td>
</tr>
</tbody>
</table>
Table 5
Performance measures for CO simulations

<table>
<thead>
<tr>
<th>No. of station(s) to be eliminated</th>
<th>Optimal station(s) combination that can be eliminated</th>
<th>Interpol. method</th>
<th>RMSE</th>
<th>$r^2$</th>
<th>MAPE</th>
<th>NSE</th>
<th>ACFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>IDW</td>
<td>0.144</td>
<td>0.877</td>
<td>6.608</td>
<td>0.841</td>
<td>0.914</td>
</tr>
<tr>
<td>2</td>
<td>10, 16</td>
<td>UK</td>
<td>0.174</td>
<td>0.838</td>
<td>6.842</td>
<td>0.775</td>
<td>0.901</td>
</tr>
<tr>
<td>3</td>
<td>2, 10, 16</td>
<td>UK</td>
<td>0.237</td>
<td>0.807</td>
<td>11.283</td>
<td>0.671</td>
<td>0.863</td>
</tr>
<tr>
<td>4</td>
<td>2, 10, 14, 16</td>
<td>UK</td>
<td>0.239</td>
<td>0.718</td>
<td>16.928</td>
<td>0.659</td>
<td>0.915</td>
</tr>
<tr>
<td>5</td>
<td>2, 9, 10, 14, 16</td>
<td>IDW</td>
<td>0.236</td>
<td>0.665</td>
<td>25.182</td>
<td>0.647</td>
<td>0.953</td>
</tr>
<tr>
<td>6</td>
<td>2, 4, 9, 10, 14, 16</td>
<td>IDW</td>
<td>0.287</td>
<td>0.581</td>
<td>34.782</td>
<td>0.476</td>
<td>0.870</td>
</tr>
<tr>
<td>7</td>
<td>2, 4, 7, 9, 10, 14, 16</td>
<td>IDW</td>
<td>0.288</td>
<td>0.509</td>
<td>44.646</td>
<td>0.467</td>
<td>0.951</td>
</tr>
<tr>
<td>8</td>
<td>2, 4, 7, 9, 10, 14, 15, 16</td>
<td>UK</td>
<td>0.329</td>
<td>0.444</td>
<td>51.856</td>
<td>0.374</td>
<td>0.926</td>
</tr>
</tbody>
</table>

3.3 SO$_2$

Station 1, which is located in the outskirts of the city, was the best station to be eliminated with RMSE=2.155, MAPE=2.155 and $r^2=0.944$. For eight stations, these parameters increased to 9.962, 69.931 and 0.301, respectively, as shown in Table 4. To limit the MAPE value to <25, a maximum of five stations could be eliminated. Figure 8 shows the rest of the 11 stations required to produce the SO$_2$ concentration maps. The higher number of stations is needed for this, as the sources of SO$_2$ are distributed with a wide variation in the concentration. The stations in the
south—2, 3, 7, 8—are predominantly located in the industrial area where power plants and refineries are located. Other stations are in the residential area where exhaust from automobiles could be the source of SO$_2$. Station 9 is in the agriculture area and 14 and 16 are located in the automobile workshops and small scale industries zones.

### 3.4 CO

The results of simulations run on CO data are shown in Table 5. Since the concentrations of CO were very small, RMSE values were also very small, ranging from 0.144 to 0.329 (Figure 3).

MAPE values ranged from 6.608 to 51.586 for one station to eight station elimination as shown in Figure 4. Station 10 was the best single station that could be eliminated and the stations 2, 4, 7, 9, 10, 14, 15 and 16 were the best eight stations. However, in order for the MAPE to be within 25, a maximum of only five stations could be eliminated as shown in Table 5. The NSE, ACFT and $r^2$ values were satisfactory for this 5-station elimination. Figure 9 shows the rest of the 12 stations that are needed to make good CO concentration maps. As shown in the figure, most of the stations are located in industrial zones or the highly densely populated areas.

**Figure 9:** Stations required to generate CO concentration maps (RMSE<0.24 and MAPE<25)

<table>
<thead>
<tr>
<th>Stations needed to achieve MAPE &lt;25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pollutant</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>1, 2, 3, 5, 6, 7, 9, 10, 14, 16</td>
</tr>
</tbody>
</table>

**Table 7**

Parameters with priority as O$_3$

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>RMSE</th>
<th>$r^2$</th>
<th>MAPE</th>
<th>NSE</th>
<th>ACFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>O$_3$</td>
<td>6.114</td>
<td>0.831</td>
<td>25.322</td>
<td>0.821</td>
<td>1.016</td>
</tr>
<tr>
<td>NO$_x$</td>
<td>24.173</td>
<td>0.208</td>
<td>57.374</td>
<td>0.189</td>
<td>1.122</td>
</tr>
<tr>
<td>CO</td>
<td>0.490</td>
<td>0.234</td>
<td>82.670</td>
<td>0.221</td>
<td>1.112</td>
</tr>
<tr>
<td>SO$_2$</td>
<td>17.579</td>
<td>0.056</td>
<td>240.894</td>
<td>-4.570</td>
<td>2.302</td>
</tr>
</tbody>
</table>
3.5 Overall

As observed from Tables 2, 3, 4 and 5, there is no single station that is common among the list of possible elimination stations. Table 6 outlines the stations required against each pollutant to achieve a MAPE level of 25. This indicates that the sources of the pollutants are highly varied with respect to the location. It will be up to the decision maker to prioritize the pollutant and select the stations. Tables 7, 8, 9 and 10 illustrate the statistical parameters taking the priority stations for O\textsubscript{3}, NO\textsubscript{x}, SO\textsubscript{2} and CO, respectively. For the case of O\textsubscript{3} as priority, MAPE for NO\textsubscript{x} and CO is about 57 and 82, while SO\textsubscript{2} has a very high MAPE value of over 240. Considering NO\textsubscript{x} as priority, MAPE value for CO and O\textsubscript{3} were about 73 and 85. In this case, SO\textsubscript{2} also has a very high MAPE (over 198) as shown in Table 8. The MAPE value of O\textsubscript{3} and NO\textsubscript{x} were less than 50 in the case of SO\textsubscript{2} priority stations, while CO was a little over 80. Taking CO as priority produced a MAPE value of 31 and 60 for NO\textsubscript{x} and O\textsubscript{3} while SO\textsubscript{2} produced an exorbitantly high value of 327 (Table 10).

4. Conclusions

A simple method of optimizing the AQMN is proposed using GIS, interpolation techniques and historical data. Existing air quality stations are systematically eliminated, and the missing data is filled in using the most appropriate interpolation technique. The interpolated data is then compared with the observed data. Pre-defined performance measures RMSE, MAPE and \( r^2 \) were used to check the accuracy of the interpolated data. NSE and ACFT supported the validity of the interpolated data. The process was simulated for several sets of observed data using an algorithm developed in GIS environment. In order to achieve a MAPE value of 25 or less, no combination of stations could be eliminated for all the pollutants. The pollutants could be prioritized to achieve the most optimal scenario. The results of the prioritization showed that the most optimal scenario is for the SO\textsubscript{2} stations, which achieved MAPE for O\textsubscript{3}, NO\textsubscript{x} and CO about 28, 51 and 80, respectively.

This methodology proves to be useful to the decision makers to find optimal numbers of stations that are needed without compromising the coverage of the concentrations across the study area. Although it is a simple procedure, it does have a few limitations. A continuous set of data is required to get reliable simulation results. Due to the unavailability of such a continuous dataset, the staggered dataset is averaged as hourly data for a day and simulated in the present case study. Secondly, the process is computing intensive, requiring large computing resources, although such resources are not very expensive these days. Lastly, more parameters can be included in the performance measures to get the most appropriate results.

Acknowledgements

We gratefully acknowledge the financial support of King Abdulaziz City for Science and Technology (KACST) under grant number 32-594.
References


Cressie NAC (1993), Statistics for spatial data. New York: Wiley


ESRI, Redlands, USA (http://www.esri.com)


pollution and mortality in California”, *American Journal of Respiratory and Critical Care Medicine*, 188(5), 593-599


Koda M, Seinfeld JH (1978), Air Monitoring Siting by Objective. EPA-600/4-7-036. U.S. Environmental Protection Agency, Las Vegas, Nevada


Silva C, Quiroz A (2003), “Optimization of the atmospheric pollution monitoring network at Santiago de Chile”, *Atmospheric Environment* 37, 2337-2345

Singh KP, Gupta S, Kumar A, Shukla SP (2012, “Linear and nonlinear modeling approaches...
for urban air quality prediction”, *Science of the Total Environment* 426, 244-255


Son, J-Y, Bell ML, Lee J-T (2010), "Individual exposure to air pollution and lung function in Korea: Spatial analysis using multiple exposure approaches”, *Environmental Research*, 110, 739-749


