ESTIMATION OF SPECIATED PM$_{2.5}$ VALUES
FROM THE CHEMICAL SPECIATION NETWORK
USING THE BAYESIAN MAXIMUM ENTROPY METHOD

Stephen Teet

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of requirements for the degree of Master of Science in the Department of Environmental Science and Engineering.

Chapel Hill
2012

Approved by:

David Leith, ScD
Marc Serre, PhD
Lisa Baxter, ScD
ABSTRACT

STEPHEN TEET: Estimation of Speciated PM$_{2.5}$ Values from the Chemical Speciation Network Using the Bayesian Maximum Entropy Method (Under the direction of David Leith)

Fine particulate air pollution is of particular concern to health and mortality. Until recent years, only general PM$_{2.5}$ data were available; but with the inception of the Chemical Speciation Network, interest has mounted regarding the effects of individual components of PM$_{2.5}$. However, due to geographic and temporal scarcity of available data, researchers have been able to conclude little about the effects of these pollutants. The use of a Bayesian Maximum Entropy model could address gaps in data and lead to better analysis of specific effects of PM$_{2.5}$ components.

Data for three different pollutants were collected from the Chemical Speciation Network for seven sites in California over a period spanning five years. These data were entered into a Bayesian Maximum Entropy model to check the validity of estimations made by this model. The three models output a range of estimations, both good and bad, but overall show potential for future use.
# TABLE OF CONTENTS

LIST OF TABLES........................................................................................................ iv  
LIST OF FIGURES....................................................................................................... v
LIST OF ABBREVIATIONS........................................................................................... vi

CHAPTER

  BACKGROUND........................................................................................................... 1
  METHODS............................................................................................................... 5
  RESULTS.................................................................................................................. 11
  DISCUSSION............................................................................................................ 20
  CONCLUSIONS......................................................................................................... 23
  APPENDIX A-Relevance to public health............................................................... 24
  APPENDIX B-Data CD Contents........................................................................... 25
  REFERENCES.......................................................................................................... 26
LIST OF TABLES

Table

Table 1. Pollutants mean, standard deviation and sill values......................... 8
Table 2. Validation and training data sets counts and statistics....................... 9
Table 3. Values used for BME GUI covariance models.................................14
Table 4. Measured and estimated mean values for pollutants with $r^2$.............. 14
LIST OF FIGURES

Figure

Figure 1. Map of California CSN Sites .......................................................... 6
Figure 2. Map of mean pollutant values for CSN sites................................. 7
Figure 3. Spatial covariance models for pollutants ..................................... 12
Figure 4. Temporal covariance models for pollutants ......................... 13
Figure 5. Log-log plot of Arsenic ................................................................. 15
Figure 6. Error Map for Arsenic ................................................................. 16
Figure 7. Log-log plot of Sulfur ................................................................. 17
Figure 8. Error Map for Sulfur ................................................................. 18
Figure 9. Log-log plot of Titanium .......................................................... 19
Figure 10. Error Map for Titanium .......................................................... 20
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>As</td>
<td>Arsenic</td>
</tr>
<tr>
<td>S</td>
<td>Sulfur</td>
</tr>
<tr>
<td>Ti</td>
<td>Titanium</td>
</tr>
<tr>
<td>BME</td>
<td>Bayesian Maximum Entropy</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>CSN</td>
<td>Chemical Speciation Network</td>
</tr>
<tr>
<td>PM</td>
<td>Particulate Matter</td>
</tr>
<tr>
<td>PM$_{2.5}$</td>
<td>Fine Particulate Matter (2.5 microns or less)</td>
</tr>
<tr>
<td>µg/m$^3$</td>
<td>Micro-grams per cubic meter</td>
</tr>
<tr>
<td>ng/m$^3$</td>
<td>Nano-grams per cubic meter</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>S/TRF</td>
<td>Space/Time Random Field</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability density function</td>
</tr>
<tr>
<td>LUR</td>
<td>Land Use Regression</td>
</tr>
<tr>
<td>CMAQ</td>
<td>Community Multi-scale Air Quality model</td>
</tr>
</tbody>
</table>
CHAPTER 1
BACKGROUND

Fine particulate matter air pollution ($\text{PM}_{2.5}$) is defined as airborne particles with an aerodynamic diameter of 2.5 $\mu$m or less. $\text{PM}_{2.5}$ makes up the respirable fraction of particulate air pollution, meaning that it can be inhaled and penetrate into the gas exchange regions of the lungs (Pope and Dockery, 2006). Due to the small size and light weight of fine particulate air pollution, it tends to stay in the atmosphere for a relatively long period of time. $\text{PM}_{2.5}$ is transported over large areas and transmits indoors more easily than larger PM.

Fine particulate matter has been shown in numerous studies to have adverse effects on human health. The effects of exposure to $\text{PM}_{2.5}$ range from mild respiratory problems to respiratory illness in children and premature mortality in the sick and elderly (Pope and Dockery, 2006). In recent decades, research in epidemiology has attempted to determine specific causes of mortality with relation to increased $\text{PM}_{10}$ and $\text{PM}_{2.5}$ exposure (Franklin et al., 2006, Braga et al., 2001). Most of this previous research has focused on the impact of exposure to overall particulate pollution levels without knowing the components of PM that may directly affect specific health problems (Bell et al., 2007).
Research has shown that there is a need to better understand the spatial and temporal variability of PM$_{2.5}$ and its constituents in health effects studies (Bell et al., 2007). Since concentrations of the different PM$_{2.5}$ components often do not co-vary with PM$_{2.5}$ levels, typically making up a different percent of the sum seasonally and regionally due to source, PM$_{2.5}$ levels cannot simply be broken up into invariant fractions for study. Each constituent, or at least better defined groups of co-varying constituents, must be further scrutinized to better define health effects.

To better understand the role of specific pollutants on health, concentration measurements must be both accurate and sufficiently frequent (temporally and spatially) for use in epidemiological research. To accomplish this, the first 13 Chemical Speciation Network (CSN) monitors came online between February and July of 2000 and expanded to 54 monitoring sites by 2004. According to the U.S. EPA, the goals of the network were fourfold: to assess emission reduction efforts, to support existing modeling efforts, to support environmental welfare programs, and to aid in exposure and epidemiological research (Rice et al, 2004).

Health effects research dealing with specific fine particulate pollutants requires several important pieces. Spatially, data need to be collected within a reasonable distance of the study area to be valid for the research. Good temporal resolution is also crucial because 1) there are often lag times between pollution spikes and health effects and 2) the duration of exposure is very important to the outcome. For these reasons, pollution concentration values are needed daily for good statistical precision (Bell et al, 2007).

In the present study, a subset of three specific pollutants was chosen for analysis. Based on data availability and differences in sources, PM$_{2.5}$ concentrations of arsenic,
sulfur and titanium were used in the predictive model. These elements stand as proxies for the feasibility of predicting speciated PM$_{2.5}$ values using the presented method.

Arsenic (As) is a common metal element found throughout the earth’s crust. Thus, it is a common particulate pollutant that results from the smelting of various non-ferrous ores, especially copper, as well as from other industrial activities including coal burning. As is a highly toxic mutagen. Exposure to high levels causes acute health problems; prolonged exposure, even to slightly elevated levels, has been shown to cause cancer, particularly in the lungs (Hayes, 1997).

Sulfur (S) particulate matter, a carcinogen, is a common byproduct of burning diesel fuel (S compounds being the most abundant non-hydrocarbons in petroleum) and coal (Liang et al., 2006). Currently, efforts to reduce pollution from S and S oxides have increased the use of reduced S diesel fuels in developed countries. Although S is a common byproduct in all coal burning operations, the source of the coal is important to the levels of S produced. Coal mined in the eastern and southeastern United States has substantially higher levels of S than that mined in the west. Because most coal is used regionally, S pollution from coal burning is less prevalent in the western states.

Titanium (Ti) has an inflammatory effect on the lungs when inhaled and damages the membranes of red blood cells. Ti is relatively uncommon and is found in the atmosphere as a result of petroleum combustion (Gilmour et al, 1997) as well as from crustal sources.

Modern spatiotemporal geostatistics are the major method used to integrate natural data into an analysis framework. These methods can be summarized as probability models that are derived at some structural stage using a maximally
informative rule that satisfies general constraints, and then are updated at an integration stage that merges an array of site-specific data using a conditionalization rule (Christakos, 2002). The Bayesian/Maximum Entropy (BME) method, a type of spatiotemporal geostatistics, is a recent approach that deals with the temporal aspect of natural data in geographic information systems (GIS). The purpose of temporal GIS (TGIS) is to better characterize different phenomena (natural and otherwise) as well as to build a framework for prediction (Christakos et al., 2002). The development of the BME graphical user interface (BME-GUI, http://www.unc.edu/depts/case/BMElab/index.htm, Y. Akita) not only provides a less complex system for using the BME method for modeling spatiotemporal data, but also acts as a bridge to the GIS realm, making output readily available for mapping.

Daily speciated PM$_{2.5}$ data are critical for accurate epidemiology research, but due to the cost of sample collection and analysis, these data are not available in many cities. If data continue to be collected in the current manner however, it may be possible to use a space/time estimating BME model to fill the gaps in the data. The motivation for the present research is to examine the viability of filling both temporal and spatial gaps in CSN data using the BME method. Temporally, a BME model can be used to fill in gaps on a similar scale to that of the available data; since CSN data are taken every three to six days, it is possible to estimate daily values. Spatially, a BME model fills in a continuous, smoothed grid of values between data points and for a surrounding area that scales with the area covered by the sites. This procedure would also allow predictions of pollution concentrations in urban and rural areas where speciated data are not available.
For the present study, the focus is on the temporal aspect of the BME model. The first objective for determining the viability of BME GUI for estimating needed PM$_{2.5}$ pollutant values will be to determine if there is autocorrelation, spatial and/or temporal, for the data present in this study. Next, are the models able to make independent geostatistical estimations, spatially and/or temporally, with unique $R^2$ values for each pollutant? Finally, the overall usefulness of the BME GUI software and method will need to be assessed. A limitation of this study is that it does not fully consider the spatial viability of the model or BME GUI.

**METHODS**

**CSN Data Collection**

Speciated PM$_{2.5}$ data are collected by state, local or tribal agencies with funding from the U.S. EPA. Samples are collected on special Teflon, nylon and quartz filters at multiple CSN sites. The device that houses the filters has a size selection port that allows only PM$_{2.5}$ to the filters. On a regular basis, typically every three or six days, the filters are gathered and replaced with new ones and sent to the Research Triangle Institute International (RTI) for analysis. There the filters are analyzed for ions and carbon using gravimetric mass measurements, energy dispersive X-ray fluorescence, ion chromatography and a thermal optical transmittance method (Rice et al., 2004).

For the present study, the seven CSN monitors in California that collected speciation data from 2001 through 2005 were chosen (Figure 1). These monitors are a small subset of a larger, 49 site study for exposure and epidemiology analysis at the U.S. EPA. These data were downloaded from the Air Quality System Data Mart at the U.S.
EPA on February 26, 2009 (http://www.epa.gov/ttn/airs/). Using SAS software (SAS Industries, Carey, NC), the data were sorted by date and location, merged into a single data set, and unwanted and non-valid data, such as null values and extreme outliers, were removed. The data were then analyzed using BME GUI.

![Map of California showing the location of the CSN sites used in this study.](image)

**Figure 1.** Map of California showing the location of the CSN sites used in this study.

**Data Description**

The overall PM$_{2.5}$ values for the study area ranged between 13 and 26 µg/m$^3$ (Figure 2). While some pollutants showed spatial patterns in concentration levels, these patterns were not consistent across pollutants. The three pollutants of interest showed no regular spatial patterns across the study area since they have different sources.
Figure 2. Mean values for a) PM$_{2.5}$ ($\mu$g/m$^3$), b) As (ng/m$^3$), c) S ($\mu$g/m$^3$) and d) Ti ($\mu$g/m$^3$) over the study area.

BME Analysis:

A spatiotemporal analysis was run on the data by species using BME-GUI for 29 pollutant species to determine their practicality for further use. Gaseous pollutants were eliminated from the pool because more complete data are usually available for gaseous pollutants than for PM, and because of the physical differences between gaseous and particulate pollutants. From the remaining species, 17 elements were chosen based on
parameters that included a comparison of error-maps output by BME GUI and their means, standard deviations, and coefficients of variation (Table 1). From this group of 17 pollutants, three were randomly selected for further analysis: arsenic, sulfur and titanium.

Table 1. Pollutants tested for use in BME GUI with means, standard deviations (all in \( \mu g/m^3 \)), and coefficient of variation.

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Mean</th>
<th>Std.</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aluminum</td>
<td>0.0373</td>
<td>0.0809</td>
<td>2.1689</td>
</tr>
<tr>
<td>Arsenic</td>
<td>0.0016</td>
<td>0.0019</td>
<td>1.1885</td>
</tr>
<tr>
<td>Bromine</td>
<td>0.0033</td>
<td>0.0031</td>
<td>0.9518</td>
</tr>
<tr>
<td>Calcium</td>
<td>0.0713</td>
<td>0.0952</td>
<td>1.3356</td>
</tr>
<tr>
<td>Chlorine</td>
<td>0.0419</td>
<td>0.1374</td>
<td>3.2804</td>
</tr>
<tr>
<td>Copper</td>
<td>0.0052</td>
<td>0.0102</td>
<td>0.9919</td>
</tr>
<tr>
<td>Iron</td>
<td>0.0029</td>
<td>0.017</td>
<td>0.7310</td>
</tr>
<tr>
<td>Mercury</td>
<td>0.0024</td>
<td>0.017</td>
<td>0.7310</td>
</tr>
<tr>
<td>Potassium</td>
<td>0.0830</td>
<td>0.2109</td>
<td>2.5398</td>
</tr>
<tr>
<td>Manganese</td>
<td>0.0034</td>
<td>0.0061</td>
<td>1.7748</td>
</tr>
<tr>
<td>Sodium</td>
<td>0.0995</td>
<td>0.1363</td>
<td>1.3700</td>
</tr>
<tr>
<td>Lead</td>
<td>0.0053</td>
<td>0.0104</td>
<td>1.9805</td>
</tr>
<tr>
<td>Sulfur</td>
<td>0.0015</td>
<td>0.0022</td>
<td>1.4804</td>
</tr>
<tr>
<td>Selenium</td>
<td>0.0067</td>
<td>0.0085</td>
<td>1.2751</td>
</tr>
<tr>
<td>Titanium</td>
<td>0.0028</td>
<td>0.0038</td>
<td>1.3675</td>
</tr>
<tr>
<td>Vanadium</td>
<td>0.0168</td>
<td>0.0393</td>
<td>2.3438</td>
</tr>
</tbody>
</table>

Using SAS, data for the three pollutant species of interest were checked for missing and null values as well as outliers (95% tolerance) over the study area; these days were eliminated from the data set. Next, twenty-five days having data available at multiple sites were removed to use as a validation data set. The pollutant data for the remaining 610 measurement days, the training set, were then formatted for use in BME-GUI. The validation data were chosen at semi-equal intervals for the study period,
possibly introducing bias into the estimation. This can be seen in differences in mean values and standard deviation of validation versus training data for the pollutants (Table 2).

### Table 2

Number of days and values for validation and training data sets. Means and standard deviations of both data sets are provided for the three pollutants (values in $\mu g/m^3$).

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Data</th>
<th>Days</th>
<th>Values</th>
<th>Mean</th>
<th>Std.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arsenic</td>
<td>Validation</td>
<td>25</td>
<td>138</td>
<td>0.00167</td>
<td>0.00403</td>
</tr>
<tr>
<td></td>
<td>Training</td>
<td>610</td>
<td>3097</td>
<td>0.00152</td>
<td>0.00108</td>
</tr>
<tr>
<td>Sulfur</td>
<td>Validation</td>
<td>25</td>
<td>133</td>
<td>0.639</td>
<td>0.569</td>
</tr>
<tr>
<td></td>
<td>Training</td>
<td>610</td>
<td>3101</td>
<td>0.660</td>
<td>0.516</td>
</tr>
<tr>
<td>Titanium</td>
<td>Validation</td>
<td>25</td>
<td>138</td>
<td>0.00764</td>
<td>0.00629</td>
</tr>
<tr>
<td></td>
<td>Training</td>
<td>610</td>
<td>3097</td>
<td>0.00832</td>
<td>0.00759</td>
</tr>
</tbody>
</table>

Training data for each of the three pollutants species studied were individually entered into BME GUI to determine the covariance function. The concentration data are plotted in the software by the Space/Time Random Field (S/TRF), $X(p)$. In this function, $p$ represents the location of each measurement in space and time. The pollutants values are then modeled to give a collection of possible fields, $\mathcal{X}(p)$. The S/TRF is written as a set of random variables at given points in space and time:

$$x_{\text{map}} = (x_1, \ldots, x_v) \text{ at } p_{\text{map}} = (p_1, \ldots, p_v)$$

where $x_i$ ($i = 1 \ldots v$) are random variables at space/time locations $p_i$, and the vector $x_{\text{map}}$ the collection of random variables $(x_1, \ldots, x_v)$.

A precise framework is provided by the BME method to process the possible knowledge base $K$ that typifies the PM$_{2.5}$ constituent across the study area for the study
period January 1, 2001 through December 31, 2005. This knowledge base can be broken up into two parts: the general knowledge base, \( G \), and the knowledge base specific to the site, \( S \). The general knowledge base consists of global characteristic of the S/TRF including the mean trend, \( m_x(p) = E[X(p)] \) and the covariance, \( c_x(p, p') = E[(X(p) - m_x(p))(X(p') - m_x(p'))] \). \( E[\] \) represents the expectation operator as defined by the probability density function (PDF) of the given S/TRF (Akita et al., 2007).

For the space/time domain being mapped, \( S \) represents the measured site-specific CSN data. The general knowledge base, \( G \), is examined to obtain the prior PDF describing the S/TRF for the PM species of interest. The prior PDF is updated using Bayesian conditionalization on the site-specific knowledge base \( S \), leading to a posterior PDF describing the PM species concentrations at any space/time estimation point in the study area. The posterior PDF combines all knowledge bases creating \( K = G \cup S \), and provides a complete stochastic description of the specie concentration to estimate (Akita et al, 2007).

For this study, the log-values were used in BME-GUI to create the models for estimating data. To obtain estimated values, the mean values raster files output by BME GUI were first imported into ArcGIS (Esri Inc., Redlands, CA). This software was then used to determine the estimated mean values for each site on each day that was removed from the dataset for testing. These estimated data were converted from log to actual values, compared to the validation data set, and used for further statistical analysis.

Error maps were also created using the variance raster files output by BME GUI. The variance calculated in BME GUI represents the probable error of the estimated mean values for the given parameter. For a given measurement, the variance for that site on the
same day should approach zero as this value is included in the theoretical BME framework. The further away in space and time an estimate is made, however, the higher the error will be; this higher error can be seen in higher variance values in the raster maps output by BME GUI.

RESULTS

Values for each pollutant were plotted within BME GUI considering the S/TRF by \( X(p) \) at \( cx(r, \tau = 0) \) and \( cx(r = 0, \tau) \) to produce a full covariance function. The spatial and temporal covariance models, represented in Figures 3 and 4, respectively, are used by BME GUI to determine the effect of the data points on the model with increasing distance and time. The vertical axes (covariance) represent the degree of correlation between the model and the data. The horizontal axes represent the lag (either distance or time) over which this correlation diminishes. Using these experimental covariance models (dots), theoretical best-fit covariance models (curves) were produced. This process gives a best-fit, non-separable covariance model comprised of two exponential components:

\[
(cx(r, \tau) = c_1 \exp \left( \frac{-3r}{a_{r1}} \right) \exp \left( \frac{-3r}{a_{r1}} \right) + c_2 \exp \left( \frac{-3r}{a_{r2}} \right) \exp \left( \frac{-3r}{a_{r2}} \right),
\]

where \( c \) represents the sill values, which represent the variance and are equal to the square of the standard deviation; \( a_r \) is the spatial range; and \( a_t \) is the temporal range sill for the covariance model. For this study, the best-fit values for the presented model are given in Table 3. After these models were determined and input into BME GUI, the S/TRFs were calculated and the mean estimation rasters were calculated and output.
Figure 3. Spatial covariance model for a) Arsenic, b) Sulfur and c) Titanium. One degree is equal to 112 km.
Figure 4. Temporal covariance model for a) Arsenic, b) Sulfur and c) Titanium.
Table 3. Values used for pollutants covariance models. All space \((\alpha r)\) values are in decimal degrees and all time \((\alpha \tau)\) values are days.

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>(c_1)</th>
<th>(\alpha_{r1})</th>
<th>(\alpha_{r2})</th>
<th>(c_2)</th>
<th>(\alpha_{t1})</th>
<th>(\alpha_{t2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arsenic</td>
<td>0.248</td>
<td>5</td>
<td>20</td>
<td>0.1</td>
<td>10</td>
<td>250</td>
</tr>
<tr>
<td>Sulfur</td>
<td>0.5</td>
<td>7.5</td>
<td>50</td>
<td>0.18</td>
<td>10</td>
<td>750</td>
</tr>
<tr>
<td>Titanium</td>
<td>0.429</td>
<td>7.5</td>
<td>30</td>
<td>0.195</td>
<td>10</td>
<td>500</td>
</tr>
</tbody>
</table>

The values estimated with BME GUI ranged from one third of measured values to over five times the measured values for all three pollutants. Arsenic and titanium values were typically overestimated by BME GUI while estimated values for sulfur were more evenly distributed. A comparison of mean estimated to mean measured values maintains this observed pattern in the data, with the estimated value of \(As\) being approximately 1.27 times the measured value and that of \(Ti\) being 1.28 times its measured value (Table 4). The log-based \(R^2\) values, also shown in Table 4, show the proportion of variability between the predicted and measured models.

Table 4. Mean measured and estimated values (\(\mu g/m^3\)) and their \(R^2\) values for the three pollutant over the study area.

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Means</th>
<th>(R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimated</td>
<td>Measured</td>
</tr>
<tr>
<td>Arsenic</td>
<td>0.00212</td>
<td>0.00167</td>
</tr>
<tr>
<td>Sulfur</td>
<td>0.663</td>
<td>0.639</td>
</tr>
<tr>
<td>Titanium</td>
<td>0.00976</td>
<td>0.00764</td>
</tr>
</tbody>
</table>

Arsenic:

The arsenic values predicted by BME GUI tended to be higher than validation values, except for higher measured values (over 2.5 \(ng/m^3\)). Figure 5 shows a graphical comparison of the estimated and measured values using a log-log scale. This method was used to compare the estimated and measured values for all pollutants since log values
were used for estimation models in BME GUI. For As, the slope intercept was forced through zero, giving a slope of 0.197 with a 95% slope confidence interval of 0.110 to 0.285. The log-log calculated r-squared value for As was 0.128.

![Log-Log plot of estimated versus measured values of As with trend-line (log y = 0.197 log x + 0, r² = 0.128).](image)

**Figure 5.** Log-Log plot of estimated versus measured values of As with trend-line (log y = 0.197 log x + 0, r² = 0.128).

The BME error map for arsenic (Figure 6) shows variance for the modeled log estimation values. The variance calculated by BME ranges from 0.121 near the CSN sample sites to 0.318 on the outer limits of the model. This is the expected pattern of error in any such geostatistical model.
**Figure 6.** Map of variance for arsenic for the study region based on the BME GUI output log estimated mean and variance maps.

**Sulfur:**

The estimates of $S$ acquired with BME-GUI were fairly evenly dispersed between under- and over-estimation. Figure 7 is the log-log plot of BME estimated versus measured $S$ values. The log-log trend line was forced through zero with a slope of 0.796 and an $r^2$ value of 0.784, and the 95% slope confidence interval is 0.724 to 0.868.
Figure 7. Log-Log plot of estimated versus measured values of $S$ with trend-line ($\log y = 0.796 \log x + 0$, $r^2 = 0.784$).

The error map for sulfur (Figure 8) shows the variance estimated by BME GUI for the modeled log estimation values. The range of BME estimated variance runs from 0.113 at the CSN sites to 0.570 in the outlying regions of the model.
Figure 8. Map of variance for sulfur for the study region based on the BME GUI output log estimated mean and variance maps.

Titanium:

Figure 9 shows the BME estimated values versus the measured values $Ti$ values. Estimated titanium values tended to be higher than the validation values. The log-log trend line was forced through zero and has a 95% slope confidence interval of 0.743 to 0.999 with an $R^2$ of 0.567.
The calculated error map for titanium (Figure 10) shows the variance for the modeled log values estimated by BME GUI. The variance map for the Ti model ranges from 0.147 at CSN sites to 0.528 at the fringes of the study area estimated by the model.
**DISCUSSION**

The BME models presented in this study show promise for the use of BME GUI for the estimation of pollutant values in temporal gaps in the CSN network. This finding is seen in the autocorrelation of the pollutants, illustrated by the covariance plots (Figures 3 & 4). The differing degree of temporal autocorrelation between the pollutants ($\alpha_{1}$, Table 3), seems to directly affect the $R^2$ values of the estimates as there is no significant difference in spatial correlation ($\alpha_1$), except for in As, which also has the lowest $R^2$ value. Specifically, the greater (longer time period) the temporal autocorrelation of a pollutant, the better the estimation seems to be. The source of the component of PM$_{2.5}$ is a likely driver of temporal autocorrelation lag period due to the effect on particle size. Particles
from crustal sources such as arsenic tend to have larger aerodynamic diameters and thus settle out of the atmosphere more readily, whereas particles from combustion sources will typically have smaller aerodynamic diameters and stay suspended longer. The range of $R^2$ values for the three presented pollutants suggests a varying degree of usefulness of the model whatever the cause may be. For $S$ and $Ti$, $R^2$ values were within the range seen in a review of the literature (Ryan and LeMasters) on Land Use Regression (LUR) models used in calculation air pollution (0.54 – 0.81).

There is a potential concern with overestimation in the models that is likely caused by sampling used to create the training and validation data sets. Looking at the training versus validation data sets for $S$ and $Ti$ (Table 2), it is notable that the means of the training data were higher than those of their validation data. Similarly, the comparison of the measured (validation) data against data estimated with the BME GUI method shows that estimated values were higher for both $S$ and $Ti$. This was not the case for As, however, which had a larger mean for the validation data than the training data. Despite this, BME GUI gave overestimations for As as well, but having such a low $R^2$ value suggests that it was not as useful for estimating As. Since training data are used by BME GUI to create the model and make predictions, higher or lower training data values will skew the predicted values accordingly.

It is also possible to address spatial gaps in the CSN data using BME GUI by removing one or more site(s) data to create a spatial validation set. Based on the apparent influence of temporal lag on the results of the present study, it is reasonable to expect that the spatial lag ($\alpha_\tau$) of a given component of PM$_{2.5}$ would be predictive of the accuracy of the model. Since As has a shorter spatial lag (5) than $S$ and $Ti$, estimated values of As
would probably again have the lowest $R^2$ values when tested against the spatial validation set.

Geostatistical models, such as the BME GUI model presented here represent a good cost and time effective alternative to current options. To measure the pollutant levels at higher spatial and temporal resolutions would cost more than is already being spent on the CSN project. Other possible methods of estimations, such as the CMAQ (Community Multi-scale Air Quality model) program and LUR models are more time intensive due to the use of multiple data types, including emissions and meteorological data, and often use several programs. Since CSN data are already available, BME GUI is also more cost effective than remote sensing estimations, which is also limited in the pollutants that can be estimated (primarily gaseous) based on current methods.

The presence of BME GUI makes the BME model the ideal geostatistical model to begin estimations of air pollution. The graphical framework of BME GUI presents an interactive environment for building a BME model without the need of knowledge of a programming language. Once one has a theoretical understanding of BME and a technical knowledge of the software, it is relatively easy to build models and output desired predictions from the data.

One way to possibly improve this geostatistical model is through the use of soft data. Soft data may exist in the form of (other) physical data, be experiential, or inferred data (Christakos, 1998). The BME method provides a theoretically sound and technically operational framework in which to incorporate soft data analysis. It is possible, then, that by incorporating outside data, specifically the available daily PM$_{2.5}$ values from the CSN
and other sites, as soft data to be additionally incorporated into the estimation models, that significantly better spatiotemporal estimations of the pollutants could be obtained.

Ultimately, the BME method, like all spatiotemporal geostatistic methods, is limited by the frequency, both temporally and spatially, of available data. To obtain better estimations, either more frequent data will be necessary or some sort of proxy data, such as using the aforementioned soft data method, may be needed. In the absence of more cost and time effective options, BME GUI has potential usefulness in air pollution estimation for exposure and epidemiology studies.

CONCLUSION

Based on the present findings, it is possible to use BME GUI to obtain estimations of air pollution for some species. While such estimations may be limited based on the level of spatial and temporal autocorrelation of a given pollutant, it is viable option considering the monetary and time costs of other methods currently being used. BME GUI is a more promising method because this geostatistical model is an easy to use, interactive program.
APPENDIX A

Relevance to Public Health:

The negative effects of particulate air pollution have been studied for years. The smaller the particles, the further into the lungs they are inhaled, the harder they are to dislodge, and the more likely they are to be absorbed further into the cardiovascular system. The overall effects of PM$_{2.5}$ are reasonably understood from increased respiratory and cardiopulmonary disease to increased mortality. What is not as well understood is what portions of the particulate pollution contribute to which health concerns. While the Chemical Speciation Network now offers specific pollutant data every few days in select cities, it is not collected often enough to perform complete epidemiological studies. The successful use of a modern spatiotemporal geostatistic method, such as the BME model presented in this study, to fill in the temporal (and spatial) holes in the data would allow for more complete studies of these effects.
APPENDIX B

Contents of Data CD:

CSN (folder):
  CSN_final.sas7bdat (complete CSN dataset used for study)

As (folder):
  As.csv (complete arsenic dataset)
  As_25.csv (formatted arsenic training dataset)
  BME (folder): contains all output arsenic raster files in .asc format

S (folder):
  S.csv (complete sulfur dataset)
  S_25.csv (formatted sulfur training dataset)
  BME (folder): contains all output sulfur raster files in .asc format

Ti (folder):
  Ti.csv (complete titanium dataset)
  Ti_25.csv (formatted titanium training dataset)
  BME (folder): contains all output titanium raster files in .asc format
REFERENCES


