Application of Remotely-sensed Aerosol Optical Depth in Characterization and Forecasting of Urban Fine Particulate Matter

A dissertation presented to
the faculty of
the Russ College of Engineering and Technology of Ohio University

In partial fulfillment
of the requirements for the degree
Doctor of Philosophy

Shanique L. Grant
August 2014

© 2014 Shanique L. Grant. All Rights Reserved.
This dissertation titled

Application of Remotely-sensed Aerosol Optical Depth in Characterization and
Forecasting of Urban Fine Particulate Matter

by

SHANIQUE L. GRANT

has been approved for

the Department of Chemical and Biomolecular Engineering

and the Russ College of Engineering and Technology by

Kevin C. Crist
Professor of Chemical Engineering

Dennis Irwin
Dean, Russ College of Engineering and Technology
ABSTRACT

GRANT, SHANIQUE L., Ph.D., August 2014, Chemical Engineering

Application of Remotely-sensed Aerosol Optical Depth in Characterization and Forecasting of Urban Fine Particulate Matter

Director of Dissertation: Kevin C. Crist

Emissions from local industries, particularly coal-fired power plants, have been shown to enhance the ambient pollutant budget in the Ohio River Valley (ORV) region. One pollutant that is of interest is PM$_{2.5}$ due to its established link to respiratory illnesses, cardiopulmonary diseases and mortality. State and local agencies monitor the impact of the local point sources on the ambient concentrations at specific sites; however, the monitors do not provide satisfactory spatial coverage.

An important metric for describing ambient particulate pollution is aerosol optical depth (AOD). It is a dimensionless geo-physical product measured remotely using satellites or ground-based light detection ranging instruments. This study focused on assessing the effectiveness of using satellite aerosol optical depth (AOD) as an indicator for PM$_{2.5}$ in the ORV and two cities in Ohio. Three models, multi-linear regression (MLR), principal component analysis (PCA) – MLR and neural network, were trained using 40% of the total dataset. The outcome was later tested to minimize error and further validated with another 40% of the dataset not included in the model development phase. Furthermore, to limit the effect of seasonality, four models representing each season were created for each city using meteorological variables known to influence PM$_{2.5}$ and AOD concentration.
GIS spatial analysis tool was employed to visualize and make spatial and temporal comparisons for the ORV region. Comparable spatial distributions were observed. Regression analysis showed that the highest and lowest correlations were in the summer and winter, respectively. Seasonal decomposition methods were used to evaluate trends at local Ohio monitoring stations to identify areas most suitable for improved air quality management. Over the six years of study, Cuyahoga County maintained PM$_{2.5}$ concentrations above the national standard and in Hamilton County (Cincinnati) PM$_{2.5}$ levels ranked above the national level for more than half the study period. Therefore, forecasting models were developed for these two locations.

All models had AOD as a significant predictor variable. In Cincinnati, the neural network and MLR models were the most useful for the summer and fall seasons; while, in the neural network explained most of the observed variance in Cleveland.
To my parents, Donald and Sharon Grant...
ACKNOWLEDGMENTS

I am indebted to the extraordinary people who offered me their unwavering support and guidance throughout my PhD study and research. I would like to express my deepest gratitude to my advisor, Dr. Kevin Crist, who engenders a great atmosphere for doing research and has an inordinate amount of patience. I would not have been able to complete my dissertation without his insightful guidance.

Dr. Valerie Young has been a mentor, colleague and friend and I want to thank her for her willingness to help and for generously sharing her time and ideas. I would also like to thank my dissertation committee, Dr. Douglas Goetz, Dr. Gaurav Sinha and Dr. James Lein, for their support throughout this arduous experience.

I owe an enormous gratitude to the Chemical and Biomolecular Engineering Department for providing funding and scholarship over the past 6 years. I also want to express my appreciation to Carrie Linscott who always had an open door and a willingness to help.

The support of my family has made this journey tolerable and so I thank them for motivating and supporting all my endeavors.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>3</td>
</tr>
<tr>
<td>Dedication</td>
<td>5</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>6</td>
</tr>
<tr>
<td>List of Tables</td>
<td>9</td>
</tr>
<tr>
<td>List of Figures</td>
<td>10</td>
</tr>
<tr>
<td>Chapter 1: Introduction</td>
<td>10</td>
</tr>
<tr>
<td>Particulate Matter</td>
<td>13</td>
</tr>
<tr>
<td>Regulation and Monitoring</td>
<td>14</td>
</tr>
<tr>
<td>Remote Sensing</td>
<td>16</td>
</tr>
<tr>
<td>Aerosol Optical Depth</td>
<td>17</td>
</tr>
<tr>
<td>Instruments</td>
<td>18</td>
</tr>
<tr>
<td>MODerate Imaging Spectroradiometer (MODIS)</td>
<td>19</td>
</tr>
<tr>
<td>Chapter 2: Spatio-temporal Relationship between AOD and PM$_{2.5}$</td>
<td>21</td>
</tr>
<tr>
<td>Air Quality in the Ohio River Valley Region</td>
<td>21</td>
</tr>
<tr>
<td>AOD – PM$_{2.5}$ correlation</td>
<td>23</td>
</tr>
<tr>
<td>Methodology</td>
<td>24</td>
</tr>
<tr>
<td>Regional Data</td>
<td>24</td>
</tr>
<tr>
<td>Local Data</td>
<td>25</td>
</tr>
<tr>
<td>Trend at local sites</td>
<td>26</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>31</td>
</tr>
<tr>
<td>ORV Analysis</td>
<td>31</td>
</tr>
<tr>
<td>Air Quality in Ohio</td>
<td>34</td>
</tr>
<tr>
<td>Seasonal comparison</td>
<td>37</td>
</tr>
<tr>
<td>Air Quality at Local Monitoring Sites</td>
<td>39</td>
</tr>
<tr>
<td>Summary</td>
<td>46</td>
</tr>
<tr>
<td>Chapter 3: PM$_{2.5}$ Forecasting</td>
<td>47</td>
</tr>
<tr>
<td>Introduction</td>
<td>47</td>
</tr>
<tr>
<td>AOD and Forecasting Models</td>
<td>48</td>
</tr>
<tr>
<td>Methodology</td>
<td>50</td>
</tr>
<tr>
<td>Satellite Data</td>
<td>51</td>
</tr>
<tr>
<td>Meteorological Data</td>
<td>52</td>
</tr>
<tr>
<td>Data Collocation and Integration</td>
<td>55</td>
</tr>
<tr>
<td>Models</td>
<td>57</td>
</tr>
<tr>
<td>Multi-linear regression (MLR)</td>
<td>57</td>
</tr>
<tr>
<td>Principal component analysis and MLR</td>
<td>58</td>
</tr>
</tbody>
</table>
Neural networks & air quality................................................................. 58
Model algorithm ..................................................................................... 60
Model Evaluation ...................................................................................... 61
Results and Discussion - Forecasting ..................................................... 64
Descriptive Statistics .............................................................................. 64
Forecasting Model 1 – Multiple Linear Regression ............................... 67
   MLR- AQI classification ..................................................................... 72
Forecasting Model 2 – PCA and MLR .................................................... 74
   Cleveland - PCA .............................................................................. 74
   Cincinnati - PCA ............................................................................ 75
   Regression ....................................................................................... 75
Forecasting Model 3 - Neural Network ................................................. 81
Summary ................................................................................................. 83
Conclusion .............................................................................................. 86
References .............................................................................................. 90
Appendix A: Regional Hourly PM2.5 ..................................................... 97
Appendix B: Potential Model Input Variables ........................................ 102
Appendix C: Neural Networks and Codes .............................................. 108
**LIST OF TABLES**

Table 1: Description of active continuous (hourly) PM$_{2.5}$ monitoring sites in Central, Northeast and Northwest Ohio during the study period (2007-2012) ........................................ 28
Table 2: Description of active continuous (hourly) PM$_{2.5}$ monitoring sites in Southeast and Southwest Ohio during the study period (2007-2012) ........................................ 29
Table 3: AirNow air quality index classification (EPA, 2012) ........................................ 62
Table 4: Performance evaluation metrics used to assess modeled PM$_{2.5}$ .................. 63
Table 5: Seasonal model performance metrics for the Cincinnati - MLR .................. 69
Table 6: Seasonal model performance metrics for the Cleveland - MLR .................... 72
Table 7: Selected PC for Cleveland: B1Backward ($\lambda \geq 0.70$) and Kaiser-Guttman ($\lambda \geq 1$) ........................................................................................................................................... 76
Table 8: Selected PC for Cincinnati: B1Backward ($\lambda \geq 0.70$) and Kaiser-Guttman ($\lambda \geq 1$) ........................................................................................................................................... 77
Table 9: Seasonal model performance metrics for the Cincinnati .......................... 78
Table 10: Neural network performance statistics divided into four seasons for each city 84
Table B1: Descriptive statistics of potential variables separated into categories ........ 102
LIST OF FIGURES

Figure 1: Counties currently listed with non-attainment or maintenance designations by the EPA in the Ohio River Valley region ................................................................. 22
Figure 2: The State of Ohio divided into five regions and respective counties ........... 26
Figure 3: Distribution of active PM$_{2.5}$ continuous monitoring sites in Ohio (2007-2012) superimposed unto defined sub-regions: SE- southeast; SW – southwest; NE- northeast; NW – northwest and central ................................................................. 30
Figure 4: AOD seasonal changes and spatial distribution in the ORV from 2007-2012. 32
Figure 5: PM$_{2.5}$ seasonal changes and spatial distribution in the ORV from 2007-2012. 35
Figure 6: Seasonal correlation between MODIS AOD and PM$_{2.5}$ in the ORV. (a) – Fall; (b) – winter; (c) – spring; (d) – summer ................................................................. 36
Figure 7: Pairwise comparison showing direction and magnitude of 24-hour average PM$_{2.5}$ concentrations difference between regions. Family alpha = 0.05; Bonferroni = 0.005......................................................................................................................... 37
Figure 8: PM$_{2.5}$ seasonal contribution calculated by averaging the ranked 98th percentile observations. The positive standard error is represented by the vertical whiskers. 39
Figure 9: Trend in PM$_{2.5}$ (µg/m$^3$) monthly average concentration for northeast Ohio from 2009 to 2012. The solid red line represents the general trend and the broken line the confidence intervals. p < 0.001 = ** *; p < 0.01 = **; p < 0.05 = * and p < 0.1 = +. ..... 40
Figure 10: PolarPlot of the effects of wind speed (m/s) and direction on PM$_{2.5}$ (µg/m$^3$) at the Cuyahoga (39-035-0060) monitoring station................................................................................................................................. 42
Figure 11: PolarPlot of the effects of wind speed (m/s) and direction on PM$_{2.5}$ (µg/m$^3$) at the Cincinnati (39-061-0040) monitoring station. ......................................................................................................................... 43
Figure 12: Trend in PM$_{2.5}$ (µg/m$^3$) monthly average concentration for southwest Ohio from 2009 to 2012. The solid red line represents the general trend and the broken line the confidence intervals. p < 0.001 = ** *; p < 0.01 = **; p < 0.05 = * and p < 0.1 = +. ..... 44
Figure 13: Trend in PM$_{2.5}$ (µg/m$^3$) monthly average concentration for northwest (top) southeast (middle), bottom (central) Ohio from 2009 to 2012. The solid red line represents the general trend and the broken line the confidence intervals. p < 0.001 = ** *; p < 0.01 = **; p < 0.05 = * and p < 0.1 = +. ......................................................................................................................... 45
Figure 14: Study sites and meteorological stations in Ohio. Only those counties that are shaded contains active monitoring sites at the time of study. ........................................... 51
Figure 15: Summary plot of selected meteorological variables – Cleveland. ............... 53
Figure 16: Summary plot of selected meteorological variables – Cincinnati ................ 54
Figure 17: Time-series plots showing comparable temporal distribution at the (a) Cleveland and (b) Cincinnati observation sites. ......................................................................................................................... 56
Figure 18: Descriptive statistics and summary plot for AOD and PM$_{2.5}$ at the Cleveland study site. ................................................................. 56
Figure 19: Descriptive statistics and summary plot for AOD and PM$_{2.5}$ at the Cincinnati study site ................................................................. 57
Figure 20: Schematic of the neural network structure used in the study. ............... 61
Figure 21: Correlation matrix of significant model inputs in Cleveland. The lower panel – scatterplots with loess smoothing; diagonal - histogram of data; upper panel – Pearson’s correlation coefficient. Plotted using pairs package in r (Becker, Chambers, & Wilks, 1988). ................................................................. 65
Figure 22: Correlation matrix of significant model inputs in Cincinnati. The lower panel – scatterplots with loess smoothing; diagonal - histogram of data; upper panel – Pearson’s correlation coefficient. Plotted using pairs package in r (Becker et al., 1988). ............. 68
Figure 23: Comparison of MLR-derived PM$_{2.5}$ with observed concentration at the Cincinnati study site. The blue line in the scatter plot represents the line of best fit and the vertical shows the standard deviation. ................................................................. 69
Figure 24: Comparison of MLR-derived PM$_{2.5}$ with observed concentration at the Cleveland monitoring station. The blue line shows the line of best fit and 95% confidence interval. .............................................................................. 71
Figure 25: Graphical representation of the accuracy of the MLR models in estimating the Cleveland’s AQI categories. ................................................................. 73
Figure 26: Graphical representation of the accuracy of the MLR models in estimating Cincinnati’s AQI categories ................................................................. 73
Figure 27: Kaiser-Guttman: comparison of derived PM$_{2.5}$ with observation made at the Cincinnati monitoring station (top) | Graphical representation of the accuracy of the 3 component models in estimating Cincinnati’s AQI categories (bottom). ............................... 79
Figure 28: B1Backward: comparison of derived PM$_{2.5}$ with observation made at the Cincinnati monitoring station (top) | Graphical representation of the accuracy for 5 component model in estimating Cincinnati’s AQI categories (bottom). ......................... 80
Figure 29: 24-hour estimation of PM$_{2.5}$ modeled using neural networks. Cleveland (top row) and Cincinnati (bottom row). ......................................................... 82
Figure 30: Contingency plot of predicted 24-hour estimation of PM$_{2.5}$ (neural network). Cleveland (top) and Cincinnati (bottom). ......................................................... 85
Figure 31: Taylor diagram displaying statistical comparison the models developed for Cleveland ............................................................................................................. 86
Figure 32: Taylor diagram displaying statistical comparison the models developed for Cincinnati ................................................................................................. 87
Figure A1: Boxplots of hourly PM$_{2.5}$ concentration at monitoring sites in central Ohio. 97
Figure A2: Boxplots of hourly PM$_{2.5}$ concentration at monitoring sites in SE Ohio ........ 98
Figure A3: Boxplots of hourly PM$_{2.5}$ concentration at monitoring sites in NW Ohio ....... 98
Figure A4: PM$_{2.5}$ 24-hour 98th percentile, averaged over 3 years northeast Ohio .......... 99
Figure A5: PM25 24-hour average (98$^{th}$ percentile) in southeast Ohio.......................... 99
Figure A6: PM25 24-hour average (98$^{th}$ percentile) in central Ohio............................. 100
Figure A7: PM25 24-hour average (98$^{th}$ percentile) in northwest Ohio......................... 100
Figure A8: PM25 24-hour average (98$^{th}$ percentile) in southwestern Ohio.................. 101
Figure B1: Correlogram of potential model inputs- Cincinnati....................................... 103
Figure B2: Correlogram of potential model inputs- Cleveland......................................... 104
Figure B3: Time series plots of observed PM$_{2.5}$ and predicted PM$_{2.5}$ – 3 PCA in
Cleveland .......................................................................................................................... 105
Figure B4: Time series plots of observed PM$_{2.5}$ and predicted PM$_{2.5}$ – 5 PCA in
Cleveland, vertical bars represent ±1 standard deviation. ............................................. 105
Figure B5: PM$_{2.5}$ trends conditioned by wind direction at the Cleveland site................ 106
Figure B6: Scree plot for Cincinnati PCA selection (λ=1).................................................. 107
Figure B7: Scree plot for Cleveland PCA selection ............................................................ 107
Figure C1: Neural network used for the Cincinnati spring season.................................... 108
Figure C2: Neural network used for Cincinnati summer season....................................... 115
Figure C3: Neural network used for Cincinnati winter season......................................... 122
Figure C4: Neural network used for Cincinnati fall season............................................. 129
Figure C5: Neural network used for Cleveland fall season.............................................. 135
Figure C6: Neural network used for Cleveland winter season........................................ 142
Figure C7: Neural network used for Cleveland summer season..................................... 149
Figure C8: Neural network used for Cleveland spring season........................................ 157
CHAPTER 1: INTRODUCTION

Particulate Matter

Particulate matter (PM), or aerosols, is a general term that describes an air-suspended mixture of solid particles and liquid droplets. They are diverse and complex in their composition and physio-chemical characteristics (Kaufman, Tanré, & Boucher, 2002). PM, which is also referred to as particulate pollution, is often classified into two groups: coarse and fine. The coarse fraction, $\text{PM}_{10}$, represents particles with aerodynamic size ranging from 2.5µm to 10µm and is derived primarily from the suspension of soil, dust, volcanic ash, marine aerosols, pollen, and spores, among others. The fine particulate size fraction ($< 2.5 \ \mu m$) originates primarily from combustion of fossil fuels, but can also include transformed products such as sulfates and nitrates. Size classification of particulate matter is important as it indicates the location within the respiratory tract they will deposit. Coarse particles mainly deposit in the upper respiratory tract while fine particulates behave gas-like and are able to penetrate into the gas exchange region of the lungs (World Health Organization, 2003).

A growing body of epidemiologic evidence shows a significant association between particulate matter exposure and deteriorated human health (Atkinson, Kang, Anderson, Mills, & Walton, 2014). The severity of the impact is distinguished based on the particulate size and evidence suggests that particulate-pollution-induced health effects have a stronger correlation with $\text{PM}_{2.5}$ than it does with $\text{PM}_{10}$ (Wilson & Suh, 1997; World Health Organization, 2003). Therefore, numerous studies have focused on the effects of airborne fine PM and have concluded that exposure is associated with
respiratory illnesses in children (Dockery et al., 1989; Peters, Dockery, Heinrich, & Wichmann, 1997). Other findings indicate an increased likelihood of mortality (Davidson, Phalen, & Solomon, 2005; Ostro, 1995; Samet, Dominici, Curriero, Coursac, & Zeger, 2000), lung cancer and heart diseases (Z. Hu, 2009; Pope, Ezzati, & Dockery, 2009) as a result of long-term exposure to particulate matter. Pope et al. (2009) also found that life expectancy increased with limited exposure to fine particulate matter.

Regulation and Monitoring

The Clean Air Act (CAA) mandates the Environmental Protection Agency (EPA) to establish and regulate National Ambient Air Quality Standards (NAAQS) for six air contaminants known or suspected to be harmful to the public’s health or the environment (Clean Air Act Amendments, 1990). The EPA first included fine PM in the list of criteria pollutants in 1996 and established annual standards based on whether PM existing in the atmosphere was emitted directly from the source (primary) or particles were formed in the atmosphere following emission (secondary). The most recent particle pollution rule stipulates an annual threshold of 15µg/m³ for secondary particles and limits 24-hour average concentration to 35µg/m³ (National Ambient Air Quality Standards for Particulate Matter, 2013). These limits are calculated based on a three year rolling average of PM$_{2.5}$ concentrations at a single site or a of group sites within a region (U.S. Environmental Protection Agency, 1997).

On December 14, 2012, the US EPA promulgated new NAAQS aimed at lowering the primary annual PM$_{2.5}$ standard to 12µg/m³ and projected that by 2020, 99%
of US counties will meet this 12µg/m$^3$ mark. This standard, though deemed by the EPA sufficient falls short of the WHO’s 10µg/m$^3$ limit set in 2005.

Compliance to NAAQS depends primarily on measurements obtained from State and Local Air Monitoring Stations (SLAMS) that supply data on a 1/6-day and 1/3-day sampling schedule. In addition, these monitors are situated in urban areas or areas deemed necessary by State Implementation Plans (SIP). The resulting unbalanced and sparse distribution of monitors present a challenge for analysts and researchers since these point measurements cannot adequately characterize regional PM (Gupta et al., 2006). In addition, PM has high temporal variability and a short ambient lifetime (Charlson et al., 1992; Kaufman et al., 2002; Kaufman et al., 1997; Pinto, Lefohn, & Shadwick, 2004) necessitating the need for continuous measurements. To this end, in-situ monitors designed to acquire consecutive hourly average PM$_{2.5}$ and its precursor gases have been implemented, however they are more sparsely distributed than the bi-weekly/weekly monitoring sites.

Despite these limitations, local, state and tribal agencies have relied on the spatial and temporal data series obtained from these monitoring sites to provide data to inform policies, regulations and exposure metrics in epidemiological studies (U.S. Environmental Protection Agency, 2009; Zeger et al., 2000). Similarly, forecasting and simulations are done utilizing data obtained from these surface networks; as a result, their reliability has been hampered (Dominici, Zeger, & Samet, 2003).
Remote Sensing

Taking advantage of scientific and technological developments over the last four decades, satellites have increasingly been used to estimate atmospheric particulate content (King, Kaufman, Tanré, & Nakajima, 1999). Though early satellites were not explicitly launched with the objective of estimating aerosols, informative measurements were taken and has led to aerosol specific satellites. Lee, Li, Kim, and Kokhanovsky (2009) describe in chronological order satellites that have over the past years been used to characterize aerosols.

On-board each satellite in orbit there are sensors that are able to capture geophysical products using electromagnetic radiation (EMR) emitted and/or reflected from the earth’s surface. The EMR traveling through the atmosphere interacts with airborne aerosol particles before reaching satellite sensors and creates a distortion (Kergomard & Tanré, 1989). The quantity recorded is characterized as a measure of how much light is attenuated due to scattering and absorption and is converted using algorithms, lookup tables and mathematical models to aerosol products (Chu et al., 2002; Kaufman et al., 1997; Tanré, Kaufman, Herman, & Mattoo, 1997).

Operational remote sensing of aerosols offers an alternative that will mitigate the limitations of field data collection through complete ground coverage and frequent revisit periods derived from satellites. In general, researchers have pointed to the potential boost air quality management has, and will receive due to the capabilities of Earth Observations Systems (Hoff & Christopher, 2009; Y. Liu, Paciorek, & Koutrakis, 2009). Satellite-based measurements are inherently indirect and as such, requires “ground truth”
observations in order to acquire a comprehensive understanding of the ambient aerosol particles. The drawbacks associated with the use of remotely sensed data as a proxy for in-situ observations have been highlighted in many studies (Hoff & Christopher, 2009; Y. Liu et al., 2009; Paciorek & Liu, 2009). The most prominent constraint of satellite stand-alone data is the large number of missing retrievals caused by satellite orbit patterns, cloud cover and surface reflectivity. On the other hand, Remer et al. (2005) suggested that with daily global coverage, satellite sensors offer up the capability of delineating aerosol spatial and temporal patterns that would not be recognized otherwise due to the atmospheric short lifetime of the particles. Torres et al. (2002) concur and further claimed that the use of satellite observations is the most effective way to assess atmospheric spatiotemporal microphysics. According to Hoff and Christopher (2009) remote sensing plays an essential role in episode detections and model predictions (Hoff & Christopher, 2009).

*Aerosol Optical Depth*

One remotely-sensed product designed to describe the concentration of ambient aerosol particles is Aerosol Optical Depth (AOD, $\tau_a$). AOD, a dimensionless parameter, is a measure of attenuated light resulting from aerosol particle interaction (absorption and scattering) with radiation in the atmosphere (Kaufman et al., 2002). It is determined by the integrated extinction coefficient ($\sigma_{\text{ext}}$, the fractional depletion of radiance per unit path length (x)) over a vertical atmospheric column of unit cross section, which is a function of wavelength, $\lambda$, (Equation 1).
The accuracy and significance of the relationship between AOD and PM is dependent on the characteristics and source region of the aerosols (Kumar, Chu, Foster, Peters, & Willis, 2011). Therefore, in order to employ satellite-derived AOD as a variable for PM$_{2.5}$ estimation, local and regional influential covariates must be examined. Frequent vertical temporal mismatch between daily PM averages and one daytime snapshot of AOD also contribute to differences in the measurements (Y. Liu et al., 2009). In addition, the instrument used to record the average columnar AOD values also plays a key role in the AOD-PM relationship (Kaufman et al., 2002; Levy & Remer, 2005).

**Instruments**

Low-altitude satellites such as Terra, Aqua, Aura and Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) have recently been launched. These satellites, as part of their overall function, provide a comprehensive look at global aerosol and cloud distributions. Onboard these satellites are high-end sensors (such as, Moderate Resolution Imaging Spectroradiometer, MODIS) that provide the medium through which data is recorded.
**MODerate Imaging Spectroradiometer (MODIS)**

MODIS, flying onboard two sun-synchronous, near-polar orbiting satellites: Terra and Aqua, has been operational since December 18, 1999 and May 4, 2002, respectively. According to Remer et al. (2005), MODIS uses two algorithms designed for independent aerosol data acquisition over land and ocean. MODIS onboard Aqua and Terra provide almost complete global coverage every day, with equatorial crossing time at 10:30 AM and 1:30 PM, respectively. The repeat cycle for these two satellites is 16 days.

MODIS acquire data in 36 spectral bands across a 2330 km wide swath using a cross-track scanner (King, Kaufman, Menzel, & Tanre, 1992). Each image scene or granule is divided into 5 minute data swaths (Levy, Remer, Mattoo, Vermote, & Kaufman, 2007; Remer et al., 2005). MODIS measures the total of solar radiance scattered by the atmosphere and light reflected by the ground surface and attenuated by atmospheric transmission. The intensity of surface reflectance depends on the zenith angle, satellite position and the directional reflectance properties of land cover (King et al., 1992). The algorithm that generates AOD retrievals uses the MODIS spatial resolution of 250m (660 nm band) and 500m (470 and 550 nm bands) and produces a statistically robust product in 10 km×10 km resolution. The 470nm and 660nm are interpolated to produce AOD values at the 550 nm. MODIS AOD values range from -0.05 to 5 and its predicted uncertainty over land is expected to fall within $\Delta \tau = \pm 0.05 \pm 0.15 \tau$. Remer et al. (2005) and Levy et al. (2007) provide more in depth description of MODIS.
In this study satellite-derived geophysical products are used to describe regional and local PM$_{2.5}$ in the Ohio River Valley (ORV) region. This study will describe the retrieval and formatting of low to medium resolution MODIS AOD and examine spatial and temporal variability of AOD within the study domain. PM$_{2.5}$-AOD relationships will also be evaluated using correlations, land use and meteorological data. Finally a novel forecasting model is developed and validated for localized areas.
CHAPTER 2: SPATIO-TEMPORAL RELATIONSHIP BETWEEN AOD AND PM$_{2.5}$

Air Quality in the Ohio River Valley Region

Particulate matter (aerosols) with aerodynamic diameter less than 2.5 µm (PM$_{2.5}$) originates from both natural and anthropogenic sources. However, a major fraction of the PM$_{2.5}$ arises from chemical reactions of gaseous precursors and trace metals which are primary byproducts of fossil fuel combustion. As early as 1984 the EPA recognized the Ohio River Valley (ORV) region (Ohio, Pennsylvania, West Virginia, Kentucky, Illinois, and Indiana) as a significant source of criteria pollutants due to the highly dense distribution of coal-fired utilities (Baker, Clarke, Gerstle, Mason, & Phillips, 1984). Studies have shown that emissions from local industries, particularly coal-fired power plants have and continues to enhance the particulate pollution budget in the ORV region (M. Kim, Deshpande, & Crist, 2007; White, 2009; Wittig et al., 2004; Yatavelli et al., 2006). The six-state region (Figure 1) however, is not only a source of PM but can also be classified as a receptor zone (Anderson et al., 2004).

Particulate pollution on average has been exhibiting a general downward trend nationally (U.S. Environmental Protection Agency, 2004). For southern states, X. Hu, Waller, Lyapustin, Wang, and Liu (2013) studied this trend using satellite observations and modeled aerosol products. To our knowledge, no study has evaluated the trends of total PM$_{2.5}$ in the ORV since the implementation of the 2006 EPA amendment; consequently we examine annual changes in PM$_{2.5}$ concentration at select locations over a six year period. S. W. Kim et al. (2006) showed a downward trend in NOx emissions
(satellite-observed) in the ORV region, but since SOx is the dominant PM$_{2.5}$ component in the industrial Midwest, further analysis using total PM$_{2.5}$ is required. Furthermore, an understanding of the variations of PM both spatially and temporally is crucial, since it provides a background for making predictions about the levels of PM/AOD concentration both at the surface and above. Concerns about health and data availability also provided the impetus for this investigation. Schwartz and Dockery (1992) associated daily mortality with particulate pollution in Steubenville, Ohio; while in 2006 another study showed increased ventricular arrhythmia in the focus group and concluded that the elderly were more susceptible (Sarnat et al., 2006).

![Map of Ohio River Valley region](image)

*Figure 1*. Counties currently listed with non-attainment or maintenance designations by the EPA in the Ohio River Valley region.

There are currently 14 counties in the ORV basin that violated the 24-hour NAAQS for the 2006-2008 period and are designated nonattainment based on the 2006 PM$_{2.5}$ standard (Figure 1). Figure 1 also shows areas that had a history of nonattainment,
but are now consistently meeting the standards (maintenance areas), all of which are in northeast Ohio. While much is known about deviation from the standards, not much is known about how PM$_{2.5}$ varies spatially over time in the region. Therefore, this part of the study evaluates seasonal variations of PM$_{2.5}$ in the region and makes comparisons with satellite-derived AOD during the six year study period, 2007-2012. Further, efforts are made to examine temporal trends at local monitoring sites within Ohio as a first step towards identifying local areas that may be suitable for forecasting model training and testing.

**AOD – PM$_{2.5}$ Correlation**

Correlations between AOD and ground-based PM$_{2.5}$ have been done by numerous researchers in different parts of the world, each obtaining varying results. In the United States, low correlation (for example Pearson correlation of -0.212) has been found in the west while almost perfect correlation were identified in the east (Engel-Cox, Holloman, Coutant, & Hoff, 2004). Wang and Christopher (2003) showed correlation values ranging from 0.7 to 0.9 in Alabama while correlations ranging from 0.012 - 0.88 were found in the Pacific Northwest parts of the United States (Setton, Hystad, & Keller, 2005). Kumar et al. (2011) described likely factors that causes differences in the PM$_{2.5}$ and AOD relationship. The authors stated that aerosol sources, meteorological variables and land use characteristics are unique to regions and that extrapolation from region to region hampers PM-AOD. Therefore, each individual region or locality requires an independent study.
Methodology

Regional Data

MODIS AOD observations (Terra satellite) were retrieved from the Goddard Earth Sciences Data and Information Services Center (GES DISC) archive for 2007 to 2012. This dataset is collocated and binned to a 1 1 spatial resolution. The level 3 AOD product is merged into a single hierarchical Data Format (HDF) file and distributed as daily and monthly averages per grid. This low resolution data has reduced noise as opposed to finer resolution measurement, hence it was considered the most useful to describe the seasonal AOD spatial distributions in the region.

AirNow PM$_{2.5}$ values, derived from local and state approved method or federal reference, were downloaded from GES DISC archive. This data, just like the AOD level 3 measurements, were aggregated on 1 1 spatial resolution for North America. Since these two datasets are on the same gridded scale, one-by-one comparisons are not affected by spatial variability and mismatch. The PM$_{2.5}$ feature data was clipped to the ORV domain and interpolated within domain using ArcGIS inverse distance weighting (IDW) algorithm. A minimum of 12 points was used for the search radius and a power of 2 for the distance. Twelve seasonal maps were plotted for the study period.

The Pearson correlation coefficient (Equation 2), denoted by $r$, was used to evaluate the strength of the relationship between PM$_{2.5}$ and AOD in the ORV on a seasonal basis. Databases for each season were created for AOD and PM$_{2.5}$ and a random sample of 100 points selected from each and regressed.
\[ r = \frac{N \sum xy - (\sum x)(\sum y)}{\sqrt{[N \sum x^2 - (\sum x)^2][N \sum y^2 - (\sum y)^2]}} \]  

(2)

Where:

\( N \) = number of pairs of scores
\( \Sigma y \) = sum of \( y \) scores
\( \Sigma y \) = sum of \( x \) scores
\( \Sigma xy \) = sum of the products of pairs scores
\( \Sigma y^2 \) = sum of squared \( y \) scores
\( \Sigma x^2 \) = sum of squared \( x \) scores

**Local Data**

Continuously monitored PM\(_{2.5}\) (hourly) observations were acquired from EPA monitoring stations listed in Tables 1 and 2. Data from each station was selected if more than 85\% of the total observations remained in each year following quality control and quality assurance screening. First, flagged data listed with missing values were removed pairwise. Outlier tests were done using visual inspection of box plots and histograms. PM\(_{2.5}\) daily aggregates were made to match temporal scaling of available meteorological data.

The State of Ohio was divided into five sections as shown in Figure 2. The sections are created based on the EPA’s district offices and the counties they manage. Figure 3 shows how monitoring sites are distributed throughout the State and their active years. The southwest to northeast diagonal represents more than 80\% of the total monitoring stations in the State due to population density and industrial activities.
Trend at local sites

Trends in air quality and particulate matter in general is influenced by seasonality, thus this effect is removed from the trend analysis using seasonal trend decomposition by Loess (Cleveland, Cleveland, McRae, & Terpenning, 1990). The Theil-Sen method (Sen, 1968; Theil, 1950) is used to estimate and identify significance of trends in PM$_{2.5}$ and AOD observations. The method proceeds by comparing the difference in sequential observations ($Y_t$) to all subsequent values ($Y_{t+1}$) over time (Equation 3).

$$Slope_t = \frac{Y_{t+1} - Y_t}{(t+1) - t}$$ (3)
The magnitude of the trend is then determined by finding the median of all the changes \((slope)\) where a positive value indicates an increase of the variable over time while a negative suggests a downward trend. If there is no change the slope is zero. Statistical significance of the estimated trend is based on a 99%, 95% and 90% confidence interval and is displayed next to each trend estimator as *** \((\alpha < 0.001)\), ** \((\alpha < 0.01)\) and + \((\alpha < 0.1)\). The method is particularly useful because of its insensitivity to outliers, heteroscedasticity and missing data because it requires no assumption about distribution of the dataset. Using the Theil-Sen function within the openair R package (Carslaw & Ropkins, 2012) trends in PM\(_{2.5}\) at the local monitoring stations were determined. Trend analysis for AOD was done at two monitoring station in Hamilton and Cuyahoga county and were conditioned using local meteorological parameters to determine dependence or relationship at each site.
Table 1

*Description of active continuous (hourly) PM$_{2.5}$ monitoring sites in Central, Northeast and Northwest Ohio during the study period (2007-2012)*

<table>
<thead>
<tr>
<th>Site ID</th>
<th>County</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Central</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39-049-0028</td>
<td>Franklin</td>
<td>39.91</td>
<td>-82.96</td>
</tr>
<tr>
<td>39-049-0034</td>
<td>Franklin</td>
<td>40.0</td>
<td>-82.99</td>
</tr>
<tr>
<td>39-049-0029</td>
<td>Franklin</td>
<td>40.08</td>
<td>-82.82</td>
</tr>
<tr>
<td></td>
<td>Northeast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39-153-0017</td>
<td>Summit</td>
<td>41.06</td>
<td>-81.47</td>
</tr>
<tr>
<td>39-151-0200</td>
<td>Stark</td>
<td>40.8</td>
<td>-81.37</td>
</tr>
<tr>
<td>39-035-0060</td>
<td>Cuyahoga</td>
<td>41.49</td>
<td>-81.68</td>
</tr>
<tr>
<td>39-103-0003</td>
<td>Medina</td>
<td>41.1</td>
<td>-81.91</td>
</tr>
<tr>
<td>39-085-3002</td>
<td>Lake</td>
<td>41.72</td>
<td>-81.24</td>
</tr>
<tr>
<td>39-093-3002</td>
<td>Lorain</td>
<td>41.46</td>
<td>-82.11</td>
</tr>
<tr>
<td>39-155-0007</td>
<td>Trumbull</td>
<td>41.21</td>
<td>-80.79</td>
</tr>
<tr>
<td>39-155-0005</td>
<td>Trumbull</td>
<td>41.23</td>
<td>-80.8</td>
</tr>
<tr>
<td>39-099-0014</td>
<td>Mahoning</td>
<td>41.1</td>
<td>-80.66</td>
</tr>
<tr>
<td></td>
<td>Northwest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39-003-0009</td>
<td>Allen</td>
<td>40.77</td>
<td>-84.05</td>
</tr>
<tr>
<td>39-095-0024</td>
<td>Lucas</td>
<td>41.64</td>
<td>-83.55</td>
</tr>
</tbody>
</table>
Table 2

Description of active continuous (hourly) PM\textsubscript{2.5} monitoring sites in Southeast and Southwest Ohio during the study period (2007-2012)

<table>
<thead>
<tr>
<th>Site ID</th>
<th>County</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>39-087-0012</td>
<td>Lawrence</td>
<td>38.51</td>
<td>-82.66</td>
</tr>
<tr>
<td>39-081-0017</td>
<td>Jefferson</td>
<td>40.37</td>
<td>-80.62</td>
</tr>
<tr>
<td>39-001-0001</td>
<td>Adams</td>
<td>38.79</td>
<td>-83.53</td>
</tr>
<tr>
<td><strong>Southeast</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39-025-0022</td>
<td>Clermont</td>
<td>39.08</td>
<td>-84.14</td>
</tr>
<tr>
<td>39-061-0006</td>
<td>Hamilton</td>
<td>39.28</td>
<td>-84.37</td>
</tr>
<tr>
<td>39-061-0040</td>
<td>Hamilton</td>
<td>39.13</td>
<td>-84.5</td>
</tr>
<tr>
<td>39-061-0010</td>
<td>Hamilton</td>
<td>39.21</td>
<td>-84.69</td>
</tr>
<tr>
<td>39-113-0032</td>
<td>Montgomery</td>
<td>39.76</td>
<td>-84.19</td>
</tr>
<tr>
<td>39-165-0007</td>
<td>Warren</td>
<td>39.43</td>
<td>-84.2</td>
</tr>
<tr>
<td>39-017-1004</td>
<td>Butler</td>
<td>39.53</td>
<td>-84.39</td>
</tr>
<tr>
<td>39-135-1001</td>
<td>Preble</td>
<td>39.84</td>
<td>-84.72</td>
</tr>
<tr>
<td>39-023-0005</td>
<td>Clark</td>
<td>39.93</td>
<td>-83.81</td>
</tr>
<tr>
<td>39-057-0005</td>
<td>Greene</td>
<td>39.81</td>
<td>-83.89</td>
</tr>
<tr>
<td><strong>Southwest</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3. Distribution of active PM$_{2.5}$ continuous monitoring sites in Ohio (2007-2012) superimposed unto defined sub-regions: SE - southeast; SW – southwest; NE- northeast; NW – northwest and central.
Results and Discussion

**ORV Analysis**

For the low resolution analysis, MODIS AOD (Figure 4) was found to be in relatively good agreement with PM$_{2.5}$ measurements (Figure 5). As shown in the maps, the two quantities display similar spatial patterns. This is most prominent in the summer (June, July and August) seasons and is shown in the regression plots (Figure 6). AOD seasonal average ranged from 0.04 in winter (December, January and February) to 0.41 in the summer. Similarly, the lowest PM$_{2.5}$ average concentration (7.1µg/m$^3$) occurred in the winter, but episodic events placed the highest levels in Kentucky during the fall (September, October, November) of 2012.

General high-low concentration regions are recognized by the satellite sensor throughout the region, however this breaks down in winters. This is likely a function of the lower available number of valid satellite measurements. Invalid recording arises when large surface area is consistent and as such exhibit low contrast. Additionally, if there is snow-covered ground, the bright surface is highly reflective leading to low or invalid sensor readings. During the winter season, stable conditions, low convective forces and temperature inversions causes a build-up of pollutants near the surface. Therefore, while the ground-based monitors record high concentrations, the satellite sensors may interpret the opposite.
Figure 4. AOD seasonal changes and spatial distribution in the ORV from 2007-2012.
Precipitation also affects the AOD-PM$_{2.5}$ relationship during the spring (March, April and May) and winter. If a precipitation event is followed by a satellite overpass, aerosol loadings are likely to be missed by the satellite sensor, but recorded by the ground-based monitors. In addition, PM$_{2.5}$ observations are averaged over a period of time and satellite values represent a 5 minute snapshot of the atmospheric column. It could be argued that comparisons should only be done at the exact time of over pass and surface measurements, but that would reduce the power for any analysis since temporal matches are rare. With long term evaluations (for example, seasonal) however, the effect of temporal mismatch is expected to be reduced.

Figure 6 shows the linear regression plots for each season. Summer has the highest correlation, followed by the fall ($r =0.30$) and spring ($r= 0.25$) seasons. Zhang, Hoff, and Engel-Cox (2009) observed similar seasonal relationships for their study (in EPA regions 3, 4 and 5), except fall instead of summer had the highest correlation. In that study, the three regions included much of the east coast, Great Lakes states and extended to Florida.

From the spatial distribution (Figures 4 and 5), most of the recorded aerosols for both AOD and PM are highest north of the Ohio River in fall (except 2012) and winter, while in summer this pattern shifts to the south of the Ohio river. This change in patterns is likely a function of changes in prevailing winds. As is shown from the correlation, similar spatial patterns were not identified for AOD and PM values during the spring seasons. For AOD measurements in spring there is a distinct spatial gradient (highest aerosol loading in Indiana and northern Ohio and lowest in Kentucky and West Virginia);
the ground-based PM did not show this gradient. This variability explains the strength of
the correlation between the datasets in spring.

*Air Quality in Ohio*

Factors that influence the relationship between AOD and PM$_{2.5}$ in the ORV can
better be understood if local associations are explored. In this section, trends in PM$_{2.5}$
across the state of Ohio are examined and compared to NAAQS. This was done in order
to identify the local areas most consequential to air quality in the state and the region at
large. In addition, the results provided a criteria for selecting sites that would most benefit
from air quality management through forecasting. Analyses were done for each region
identified in Figure 2.

Figure 7 shows the comparison of PM$_{2.5}$ in each of the five regions. NE and SW
have the highest median concentration of 11.40 and 11.86µg/m$^3$, respectively.
Northwestern Ohio records the lowest levels of measured aerosols and is the only region
with median concentration below 10µg/m$^3$. This finding is not surprising as PM$_{2.5}$ point
and area sources are lowest in that part of the state. The pairwise comparison plot in
Figure 7 (right) shows the magnitude and direction of the differences between PM$_{2.5}$
concentrations in each region. The plot also displays significance, which is shown as
those differences that fall outside the –Z and +Z value zone. Regions have statistically
significant differences if their p-values are less than the Bonferonni individual estimator
($\alpha$/number of comparison).
Figure 5. PM$_{2.5}$ seasonal changes and spatial distribution in the ORV from 2007-2012.
Figure 6. Seasonal correlation between MODIS AOD and PM$_{2.5}$ in the ORV. (a) – Fall; (b) – winter; (c) – spring; (d) – summer

No county in the southeast has maintenance or non-compliant designations, however, PM$_{2.5}$ concentration levels in the region were on par with counties that have these designations. As shown in Figure 7, there is no statistically significant difference between observed PM$_{2.5}$ concentration in southeastern, central and northeast Ohio. M. Kim et al. (2007) showed that at the Athens rural site (located in southeastern Ohio), formation of secondary PM$_{2.5}$ from precursor gases, transported into the area on southwesterly winds, caused episodic spikes of PM$_{2.5}$. This theory may account for the PM$_{2.5}$ measurements that are represented as outliers in the southeast boxplot (Figure 7) and accounts for the inflated PM$_{2.5}$ hourly concentrations. The accuracy of the
southeastern estimate is also challenged because there are only three active monitoring stations to represent the 27 county region.

Figure 7. Pairwise comparison showing direction and magnitude of 24-hour average PM$_{2.5}$ concentrations difference between regions. Family alpha = 0.05; Bonferroni = 0.005.

**Seasonal comparison**

The EPA determines 24-hour compliance based on a 3 year rolling average of the annually ranked 98$^{th}$ percentile values, therefore regions that have high episodic hourly concentration may fall into compliance over the long term. For our study, we found that despite the meeting the NAAQS, seasonal changes drive the distribution or presence of ambient aerosol in the state. On average, $\sim$ 47% of the ranked 98$^{th}$ percentile PM$_{2.5}$ values occurred in the warmer months and between 2008 and 2011 more than 50% of the highest concentration occurred in summer (Figure 8). This was also the case for the entire
ORV region. In Steubenville Ohio, source apportionment studies showed summer being affected by iron/steel manufacturing, crustal matter and coal combustion (Vedantham, Landis, Olson, & Pancras, 2014).

High convective mixing during the summer as a result of higher temperatures and wind speeds lead to unstable conditions thus more dispersion. Therefore it is expected that pollutants upon emission would become diluted and be transported to distances downwind in the summers. The deviation from this general behavior is probably related to air recirculation and long range transport of particulates and precursor gases into the state, which augments and sustains the higher concentration levels. For example, Deshpande (2007) suggested that at the Athens site PM$_{2.5}$ is mainly impacted by long range transport of sulfates and is driven by meteorology. It is also likely that secondary formation of PM$_{2.5}$ as a result of accelerated chemical reactions of precursor emissions such as SOx, volatile organic compounds (VOCs) or NO contribute to the summers’ high concentrations.

Figure 8 shows that in winter there is a steady decline in PM$_{2.5}$ concentration from 2009 to 2012 and on average PM$_{2.5}$ was lowest in the winter accounting for only 13% of the total PM$_{2.5}$. Concentration levels for spring seasons showed similar trends except in 2012 where there is a sharp increase. There is no specific trend in the annual observations for fall.
Air Quality at Local Monitoring Sites

So far, generalized relationships and trends for AOD and PM$_{2.5}$ have been examined in the ORV and Ohio. Similar seasonal trends were shown for PM$_{2.5}$ in ORV and Ohio. Having determined that PM$_{2.5}$ concentrations relied on seasonal effects, further evaluations were done at local monitoring stations in order to identify areas which may benefit from forecasting models. Areas selected for model development showed no or only marginal decline in ambient particulate over the study period. Only those sites that had full 6 year records were analyzed, because for model training, testing and performance evaluation, the number of observations had to be greater than 1000 in order to obtain optimum results.

Using the Theil-Sen trend approach at each site in each region, directional (negative or positive) slopes were calculated and considered significant at either 90%, 95% or 99% confidence interval. The largest significant decrease in PM$_{2.5}$ concentration
(2.3 µg/m$^3$) was observed in Columbus, Franklin County (39-049-0034) (Figure 9) while only stations in the northeast, (Cuyahoga (39-035-0060) and Summit (39-153-0017) counties), and southwest (Clermont (39-025-023) and Warren (39-165-0007) counties) (Figure 12) showed no significant change in concentration. Even though these four stations had similar average PM$_{2.5}$ decrease of approximately 0.14µg/m$^3$, all had concentrations below the annual NAAQS threshold of 15µg/m$^3$, except in Cleveland, Cuyahoga County. As is shown in Figure 9, the trend (solid red line) is consistently above 15µg/m$^3$. This suggests that following the 2006 EPA annual standard, Cleveland is likely to at best maintain its maintenance status.

**Figure 9.** Trend in PM$_{2.5}$ (µg/m$^3$) monthly average concentration for northeast Ohio from 2009 to 2012. The solid red line represents the general trend and the broken line the confidence intervals. p $< 0.001 = **$; p $< 0.01 = *$; p $< 0.05 = *$ and p $< 0.1 = +$. 

![Graph showing PM$_{2.5}$ concentrations](image-url)
The seasonal variation of PM$_{2.5}$ at the Cleveland Cuyahoga station (located south east of the Lake Erie) conditioned by wind speed and direction is shown in Figure 10. The concentric circles show the changes in wind speed. At low wind speed (center of each plot), PM$_{2.5}$ concentrations are highest for all seasons. In summer however, there is a slight variation where high wind speed is also associated with high concentration. This can mainly be seen in the winds blowing from the west and northwest. In the spring, fall and winter months, there is a slight shift in the prevailing wind direction towards west-south-west (WSW) and high concentration fall into that wind sector. Source apportionment studies show positive linear correlation of high summer PM$_{2.5}$ with sulfates and organic carbon (Varadarajan, 2007). This implies that likely sources contributing to PM$_{2.5}$ ambient loadings in Cleveland may be attributed to industrial or fossil fuel combustion processes. The other two monitoring locations in northeastern Ohio (Figure 9) show statistically significant PM$_{2.5}$ concentration decrease; 0.41 and 1.24µg/m$^3$ in Mahoning and Lake Counties, respectively.

Site number 39-061-0040 located in Cincinnati, Hamilton County is the only other monitoring station in the state that recorded PM$_{2.5}$ higher than 15µg/m$^3$ for at least half of the study period (Figure 12). From 2007 to 2012 there was 1.23 µg/m$^3$ reduction in PM$_{2.5}$ at this Cincinnati location. Similar to the Cuyahoga monitoring station, the highest ambient particulate loading was in the summer (Figure 11). The polar plot below shows that there are episodes of high wind speed (~2m/s) that is correlated with it higher levels of PM$_{2.5}$ starting from a minimum of 13µg/m$^3$. For the other seasons, higher concentrations are associated with southerly winds. Unlike the Cleveland station, the
lowest PM$_{2.5}$ levels are in the fall instead of the winter. The northwestern, central and southeastern trend plots are shown in Figure 13. All monitoring locations saw a significant decrease in PM$_{2.5}$ and recorded levels well below annual 2006 EPA NAAQS.

*Figure 10. PolarPlot of the effects of wind speed (m/s) and direction on PM$_{2.5}$ (µg/m$^3$) at the Cuyahoga (39-035-0060) monitoring station.*
Figure 11. PolarPlot of the effects of wind speed (m/s) and direction on PM$_{2.5}$ (µg/m$^3$) at the Cincinnati (39-061-0040) monitoring station.
Figure 12. Trend in PM$_{2.5}$ ($\mu g/m^3$) monthly average concentration for southwest Ohio from 2009 to 2012. The solid red line represents the general trend and the broken line the confidence intervals. $p < 0.001 = ***$; $p < 0.01 = **$; $p < 0.05 = *$ and $p < 0.1 = +$. 
Figure 13. Trend in PM$_{2.5}$ ($\mu$g/m$^3$) monthly average concentration for northwest (top) southeast (middle), bottom (central) Ohio from 2009 to 2012. The solid red line represents the general trend and the broken line the confidence intervals. $p < 0.001 = **$; $p < 0.01 = ***$; $p < 0.05 = *$ and $p < 0.1 = +$. 
Summary

Three tiers of analyses were conducted in this chapter. First general PM$_{2.5}$-AOD time series and regression analyses were done on the ORV region which highlighted distinct seasonal patterns in both quantities. The regression analysis revealed that in summer and spring PM$_{2.5}$ and AOD had a higher association, while the colder season was not well represented. Spatial distribution throughout the ORV showed the satellite-derived AOD data in relatively good agreement with ground-based point observations. The second tier analysis focused on PM$_{2.5}$ in five regions in Ohio as a step towards determining high risk regions that would benefit from air quality forecasting. In the final level, Theil-Sen seasonal decomposition was used to ferret trends at local monitoring station and highlight significance of the magnitude of change over time. Two cities, Cincinnati and Cleveland maintained high levels of PM$_{2.5}$ throughout the study period.

Based on the analyses done, Cleveland (Cuyahoga County) and Cincinnati (Hamilton County) were selected for forecasting modeling development (Chapter 3) using AOD as a covariate. Forecasting models currently exist for these cities, but they often underperform, therefore adopting new methods of estimation is relevant. Figures 10 and 11 show the dependence of PM$_{2.5}$ on wind speed and direction, therefore historical and forecasted meteorological variables were included in the model development process. In the next chapter three models are presented and their predictive performances evaluated on a seasonal basis.
CHAPTER 3: PM$_{2.5}$ FORECASTING

Introduction

Air quality forecasting tools inform communities and the general public of unhealthy air quality levels from which they are able to limit or avoid exposure. Forecasting tools are also utilized by environmental regulatory agencies to instigate actions and alerts especially urban areas. However, the existing tools are often limited by prediction accuracies, particularly for PM$_{2.5}$. In an effort to maintain near real-time information about air quality, the EPA, National Oceanic and Atmospheric Administration (NOAA), local, state and tribal agencies joined forces to created AirNow (airnow.gov, 2005). The website reports air quality index (AQI) for ozone and PM$_{2.5}$ for over 300 cities across the United States (Al_Saadi et al., 2005 & Szykman et al., 2004).

In a study spearheaded by IDEA (Infusing satellite Data into Environmental Applications) done to improve the national air quality forecasting tool, Szykman et al. (2004) incorporated MODIS derived AOD, modeled meteorological parameters in combination with monitoring site point observations. The IDEA project sought to achieve five objectives, but while achieving the goals, gave inconclusive results regarding whether the forecasting tools’ performance changed or improved. Arising out of this initiative was the 48-hour AOD forward trajectory from areas with high AOD, 850 mb winds. This 48 hour projection may contribute to the overall model error due to diurnal variability in meteorological conditions. The Ohio University, Center for Air Quality developed a multi-linear regression forecasting tool for Dayton and Cincinnati
metropolitan areas. The model shows relatively good correlations with observations but often misses episodic days. This misclassification is influenced by weather patterns, micro-meteorology and land use characteristics.

PM$_{2.5}$ has a lifetime ranging from days to weeks depending on the dynamics of the atmosphere and is able to be transported for up to 1000 km or more from its source (Seinfeld & Pandis, 1998). PM recirculation, emphasized by local winds, advective (horizontal) and convective (vertical) conditions (primarily temperature and winds), enables dispersion and dilution within the planetary boundary layer. High surface pressure or thermal inversions are conducive to high PM$_{2.5}$ concentrations while precipitation facilitates atmospheric scavenging thus reducing PM$_{2.5}$ levels in the atmosphere. Solar radiation and surface temperature affect the chemical kinetics of PM$_{2.5}$ formation. Therefore, any improved forecasting model has to consider these effects.

**AOD and Forecasting Models**

AOD measurements, though less precise than ground-based observations reflect the aerosol concentrations in the atmosphere and offer a broad synoptic view over a wide domain. Thus, numerous studies have considered AOD as a covariate in models used to estimate ground-based PM$_{2.5}$. On a global scale, van Donkelaar et al. (2006) developed a simple relation between AOD and PM$_{2.5}$ which they modified with relative humidity. This modification is necessary to remove or minimize the hydroscopic growth effect imposed on particles in the presence of moisture. AOD is measured under ambient humid
conditions without consideration for atmospheric moisture content while PM$_{2.5}$ is quantified after the sample is heated to drive off water.

Y. Liu, Sarnat, Kilaru, Jacob, and Koutrakis (2005) also developed an empirical relationship using AOD (derived from the multi-angle imaging spectro-radiometer, MISR) combined with relative humidity and boundary layer mixing height. That study found that these atmospheric conditions significantly influence the AOD-PM$_{2.5}$ relationship. A similar study was done in Asia by Tsai, Jeng, Chu, Chen, and Chang (2011), but they corrected both PM$_{2.5}$ and AOD with relative humidity and mixing height to achieve better correlations. In this case, the inclusion of mixing height reduced the effect of convective mixing in the atmospheric column thus making the measured AOD more relatable to surface measurements. Contrarily, in Europe, Koelemeijer, Homan, and Matthijsen (2006) found that the corrective influence of relative humidity was less pronounced than correction using mixing height alone.

Goyal, Chan, and Jaiswal (2006) included temperature, solar radiation and wind speed in a regression model but showed that the meteorological variables did not improve the model capabilities. Y. Liu, Franklin, Kahn, and Koutrakis (2007) argued however, that while temperature was not a robust estimator in their model it was used because higher temperatures drive the formation of secondary particles near the surface; this leads to higher particulate concentration. Other models found in the literature (Hutchison, Smith, & Faruqui, 2005; Kumar, Chu, & Foster, 2007; Y. Liu et al., 2007; Tian & Chen, 2010) typically use the meteorological variables discussed to gain the best possible enhancement for AOD as a PM$_{2.5}$ explanatory variable.
Here, these variables (relative humidity (RH), air temperature (airTemp), surface/skin temperature (surfTemp), planetary boundary layer mixing height (MPBL)) are included in our analysis. However, we added precipitation (PRCP), visibility (VIS), wind speed (ws), horizontal (uwnd) and vertically (vwnd) resolved winds, total solar radiation (SolRad), mean sea level pressure (MSLP) and mean station pressure (MSTP) to the model development process based on their potential local influences on PM2.5 and AOD. Three predictive models: (1) Multi-linear regression, (2) Principal components and MLR and (3) Neural networks were developed and their performance evaluated based on using EPA’s forecasting metrics (Table 4).

Methodology

Cleveland (NE region) and Cincinnati (SW region) are two of the largest urban areas in Ohio. As shown in Chapter 2, these cities were located in the regions with the highest PM2.5 activity and have recurring non-attainment or maintenance statuses (based on 2006 and 1997 PM2.5 NAAQS). The trend analysis done at each of the location revealed PM2.5 levels above NAAQS annual average of 15µg/m³ for most (Cincinnati) or all (Cleveland) of the study period. Thus, they were selected as prototypes for the development of forecasting models. Figure 14 shows the study locations and weather stations used in our study.
Satellite Data

Six years (2007-2012) of MODIS level 2 AOD data were retrieved from NASA’s Goddard Earth Sciences Distribution Active Archive Center (GES DAAC) collection 5.1. The cloud-free readings from Terra’s MODIS (MOD04-L2) were selected as the primary source for AOD; however data derived from Aqua (MYD04-L2) were also acquired and used to supplement MOD04-L2 observations. Both datasets are issued on a 10 x 10 km grid.

Figure 14. Study sites and meteorological stations in Ohio. Only those counties that are shaded contains active monitoring sites at the time of study.
Simulated mixing height (km), surface temperature (K), vertical and horizontal winds (m/s) were downloaded from the Computational and Information Systems Laboratory Research Data Archive (CISL RDA). This data were modelled using inputs from observations from surface and upper air weather stations and projected onto NCEP Eta 32km/29 pressure level system (National Centers for Environmental Prediction - National Weather Service NOAA U. S. Department of Commerce, 2005). Daily meteorological information on relative humidity (%), precipitation (mm), temperature (K), visibility (m), mean sea level pressure (Pascal), wind speed (m/s), and wind direction (degrees) were acquired from the National Climatic Data Center (NCDC).

Except for principal component analysis model (discussed later), a feature selection algorithm was used to screen and prune the meteorological datasets. Variables were removed from the analyses if they had more than 70% missing or flagged values, a minimum coefficient of variation of 0.1 and minimum standard deviation of zero. Figures 15 and 16 show descriptive statistics and summary plots for the selected important variables.

For the Cleveland location, analyses revealed that the resulting important meteorological variables were relative humidity, air temperature and wind speed measured at 10m, vertically resolved wind velocity (VWND) and mixing height (Figure 15). For the Cincinnati location, variables retained were visibility, mixing height, wind speed measured at 10m, VWND and UWND (east-west horizontal winds) and precipitation (Figure 16).
McMurry, Zhang, and Lee (1996) found that relative humidity above 40% affected the aerosol optical photometric potential and suggested that in order to reduce error, relative humidity should be included when it is less 50%. Others (X. Liu et al., 2008b; Tsai et al., 2011) suggested hygroscopic growth factor correction. In this study the latter approach was chosen and the raw AOD (AODr) was modified using Equation 4 (Koelemeijer et al., 2006).

\[
AODc = \frac{AODr}{f(RH) \times MPBL}
\]  

\textit{Figure 15.} Summary plot of selected meteorological variables – Cleveland.
The hygroscopic growth factor, $f(RH)$, is defined as: $f(RH) = 1 + a(RH/100)^b$, where RH is the relative humidity (dry), ‘$a = 3.26$’ and ‘$b=3.85$’. The values for the constants ‘$a$’ and ‘$b$’ represent the mixed aerosol environment (X. Liu et al., 2008a). Most of the aerosols present in the atmosphere are within the boundary layer and as such the AOD columnar readings were also corrected with the vertical mixing height (MPBL). The boundary is considered to be well mixed.
Data Collocation and Integration

The datasets used in the study are both point (PM$_{2.5}$) and raster (AOD, 10km and NARR, 32km). Therefore to remove the misalignment, AOD rasters were resampled (maximum likelihood) to 32km grids. Using the ArcGIS spatial analyst tool “Extract values to point”, AOD and NARR data that intersected the PM$_{2.5}$ point locations or were within the assigned 10km buffer zone were extracted.

AOD measurements downloaded from MODIS onboard Terra were used as the primary dataset. However, due to low retrievals, the Terra measurements were augmented with AOD Aqua observations on days where no data were available from Terra MODIS. Paired t-test statistics were applied to the datasets to determine if there was a significant difference between Terra and Aqua AOD measurements on days when they both have valid data. The coupling of the two dataset produced 1210 observations for Cincinnati; in Cleveland 916 valid data points was the final tally.

The paired t-test analysis suggested a failure to reject the null hypothesis, indicating that there is no statistically significant difference between the mean daily Terra and Aqua AOD measurements at both study sites. These determinations were made based on p-value statistic (95% confidence level), which were found to be 0.554 and 0.456 for Cleveland and Cincinnati, respectively. These findings also indicated that using AOD data from Aqua should not significantly alter the outcome of the analysis if the dataset from both sensors were combined. Time-series plots (Figure 17) and high positive Pearson correlations, (0.90 - Cincinnati and 0.80 – Cleveland), between the AOD datasets also validate this approach.
Figure 17. Time-series plots showing comparable temporal distribution at the (a) Cleveland and (b) Cincinnati observation sites.

The PM$_{2.5}$, AOD and meteorological data summary plots comparing AOD and PM$_{2.5}$ at both sites are given in Figures 18 and 19. The plots show the number of missing data (red blocks), histograms and time-series for each dataset.

Figure 18. Descriptive statistics and summary plot for AOD and PM$_{2.5}$ at the Cleveland study site.
Figure 19. Descriptive statistics and summary plot for AOD and PM$_{2.5}$ at the Cincinnati study site

Models

Multi-linear regression (MLR)

A forward stepwise variable selection procedure was used to identify the best explanatory variables for PM$_{2.5}$ mass concentration at the selected Cincinnati and Cleveland study sites. The non-significant variables were rejected based a 95% confidence interval. The Durban-Watson statistic revealed a positive serial correlation; therefore, to minimize the effect this will have on the regression four separated models were developed. Each model represented a season (fall, spring, summer or winter). The models produced had the general form: $Y_i = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_j + C$ where $Y$ is the PM$_{2.5}$ concentration, $B_1$-$B_k$ are the coefficients corresponding to each independent variable (X) and C is the intercept.
Principal component analysis and MLR

Principal component analysis was used to summarize ground-based meteorological variables, upper atmosphere soundings and AOD into linear relationships called principal components (PCs). The PCs consist of uncorrelated variables and are ranked according to the Eigenvalues and explained variances. Identification of significant PCs to the analysis was estimated using threshold Eigenvalues and cumulative variances. PCs were identified as non-redundant components if they met the Kaiser-Guttman Criterion (Guttman, 1954; Kaiser, 1960), which suggests that only those factors with eigenvalues greater than 1 should be retained. Based on a method proposed by Jolliffe (1972), an Eigenvalue of 0.70 as the threshold can also be used to discard redundant components. This is called the B1 Backward approach. In addition to the Eigenvalue criteria, the selection is satisfied when approximately 80% of the variance is explained by the identified PCs. Both methods were used to identify the structure of regression models used to describe and predict PM$_{2.5}$. Each PC is considered to be independent in the regression.

Neural networks & air quality

Classical neural networks loosely mimic biological processes, particularly those that occur in the brain and are designed primarily to learn and recognize patterns (Gupta & Christopher, 2009). They are robust modelling tools and are not affected by the underlying data distribution or the relationship between input variables. For this reason, they have been used in cases where there are complex, nonlinear input variable
structures. As shown in Figures 15, 16, 18 through 19, the input variables are either skewed or bi-modal hence the decision to use neural networks for PM$_{2.5}$ prediction.

The application of neural networks to air quality forecasting is well documented. Comrie (1997) modeled and forecasted ozone and obtained better performance relative to standard statistical models. A network designed to predict sulfur dioxide, carbon monoxide, nitrogen dioxide and nitric oxide in Italy, included meteorological data to bolster the output (Ghazali & Ismail, 2012). In 2006, neural networks were used to forecast PM$_{10}$ in Santiago, Chile (Perez & Reyes, 2006). Gupta and Christopher (2009) and Yao and Lu (2014) developed neural network models for PM$_{2.5}$ mass concentration in southeast US and China, respectively. In both studies, remotely sensed AOD was included as a predictor variable. The China study was done to examine spatial and temporal variability of PM$_{2.5}$ while the other looked at increased accuracy of PM$_{2.5}$ prediction relative to statistical models. Those studies were concerned with large domains and as such included simulated meteorological data instead of actual measurements. In our study, predictions were made for point locations based on measured weather data that fall within a 10 km distance from the PM$_{2.5}$ study site. In addition, models were developed for each season (Appendix C) and relied only on variables that passed the feature selection algorithm. All models were developed using IBM SPSS Modeler v. 16 (2013) and statistical analysis done using R/RStudio openair (Carslaw & Ropkins, 2012) and car (Fox & Weisberg, 2010) packages.
Model algorithm

Multilayer Perceptron (MLP) neural network (NN) models were developed and trained using AOD and meteorological data as inputs and PM$_{2.5}$ as the output (target) variable. MLP is a commonly used feed-forward architecture that consists of three layer of neurons: input, hidden and output (Figure 20). As is typical for the feed-forward design, signals flow in only one direction (input to output) and all inputs are connected to all hidden neurons within the network. Each input variable is mapped to the output via weighted connections which are transferred through linear (Equation 5) and sigmoid activation (Equation 6) functions in the hidden layers.

The response variable ($Y_j$, PM$_{2.5}$) calculated from Equation 5 is compared to the target training data and the difference (error) determined. If this difference is higher than the minimum error threshold, the nonlinear activation propagates this information backwards where the weights ($w_{ij}$) associated with each input ($X_i$) is adjusted. A new response is then calculated and fed forward for a second comparison with the target. This iterative process, known as training, continues until a minima is reach or the potential of over-fitting exist. The weights indicate the strength of the relationship between the neurons and the process of weight adjustment is based on the back propagation algorithm developed by Rumelhart, Hinton, and Williams (1985).

$$Y_j = \sum_{i=1}^{n} w_{ij} X_i + w_{bj} X_b$$

(5)
\[ \sigma(Y_j) = \frac{1}{1 + e^{-Y_j}} \]  

During the training phase of the model development, 40% of the dataset was used. Model testing included another 20% of the original data while the remaining fraction was held back for validation.

![Neural network structure](image)

*Figure 20. Schematic of the neural network structure used in the study.*

**Model Evaluation**

The predicted PM$_{2.5}$ values produced from each model were used to estimate air quality index (AQI). The AQI is divided into six categories separated on a scale of 0 to 500 (Table 3). The AQI provides a means for the public to receive general air quality exposure risk. Using the predicted PM$_{2.5}$ values generated for each model the AQI values and categories were calculated. This, in addition to the metrics outlined in Table 4 was
employed to evaluate the performance of the models. Finally, a Taylor diagrams (Taylor, 2001) were used to summarized and compared the performance of the model in each city.

Table 3

*AirNow air quality index classification (EPA, 2012)*

<table>
<thead>
<tr>
<th>AQI values</th>
<th>24-hour PM$_{2.5}$ ($\mu g/m^3$)</th>
<th>AQI categories</th>
<th>Color indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 50</td>
<td>0 – 15.4</td>
<td>Good</td>
<td>Green</td>
</tr>
<tr>
<td>51 – 100</td>
<td>15.5 – 40.4</td>
<td>Moderate</td>
<td>Yellow</td>
</tr>
<tr>
<td>101 – 150</td>
<td>40.5 – 65.4</td>
<td>Unhealthy for sensitive groups</td>
<td>Orange</td>
</tr>
<tr>
<td>151 – 200</td>
<td>65.5 – 150.4</td>
<td>Unhealthy</td>
<td>Red</td>
</tr>
<tr>
<td>201 – 300</td>
<td>150.5 – 250.4</td>
<td>Very Unhealthy</td>
<td>Purple</td>
</tr>
<tr>
<td>301 – 500</td>
<td>250.5 – 500.4</td>
<td>Hazardous</td>
<td>Maroon</td>
</tr>
</tbody>
</table>
### Table 4

Performance evaluation metrics used to assess modeled PM$_{2.5}$

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction of predicted PM$_{2.5}$ within a factor of two</td>
<td>FAC2: $0.5 \leq \frac{P_i}{O_i} \leq 2.0$</td>
</tr>
<tr>
<td>Mean bias (MB)</td>
<td>$MGE = \frac{1}{n} \sum_{i=1}^{n} (P_i - O_i)$</td>
</tr>
<tr>
<td>Mean gross error (MGE)</td>
<td>$MGE = \frac{1}{n} \sum_{i=1}^{n}</td>
</tr>
<tr>
<td>Normalized mean bias (NMB)</td>
<td>$NB = \frac{\sum_{i=1}^{n} (P_i - O_i)}{\sum_{i=1}^{n} O_i}$</td>
</tr>
<tr>
<td>Normalized mean gross error (NMGE)</td>
<td>$NMGE = \frac{\sum_{i=1}^{n}</td>
</tr>
<tr>
<td>Root mean squared error (RMSE)</td>
<td>$\left( \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n} \right)^{\frac{1}{2}}$</td>
</tr>
<tr>
<td>Coefficient of Efficiency</td>
<td>$1 - \frac{\sum_{i=1}^{n}</td>
</tr>
<tr>
<td>Index of Agreement (d)</td>
<td>$d = 1 - \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (</td>
</tr>
</tbody>
</table>
Results and Discussion - Forecasting

Descriptive Statistics

The median AOD loadings (± median absolute deviation (MAD)) measured in both cities were almost identical: 0.10 ± 0.10 in Cleveland and 0.11 ± 0.10 in Cincinnati. PM$_{2.5}$ concentration levels were also not significantly different from each other, with medians of 17.8 ± 8.92 µg/m$^3$ and 16.9 ± 7.49 µg/m$^3$ for Cleveland and Cincinnati, respectively. For both PM$_{2.5}$ and AOD, the maximum readings occurred at the Cleveland location. For PM$_{2.5}$ that maximum was almost 15 units higher than the maximum fine particulate matter observed in Cincinnati (49.3µg/m$^3$).

Using the feature selection algorithm described earlier, only those variables shown in Figures 21 and 22 met the set criteria to be retained. In Cleveland those were AOD, HPBL, VWND, ws, RH and airTemp while in Cincinnati, AOD, VIS, HPBL, ws, VWND, UWND and PRCP made the cut.

The correlation between AOD and PM$_{2.5}$ was higher in Cincinnati for both corrected and uncorrected AOD values. The first obvious explanation for this could be the higher number of daily episodic PM$_{2.5}$ (outliers) in Cleveland as shown in the correlation panel scatterplot in Figure 21. However, aerosols hygroscopic growth as result of high humidity may be the main driving force. In Cleveland, the minimum RH value was 45% and only 7 of the 916 observations had RH values less than 50%. This suggested that the particle growth physical and optical properties of the aerosols may
have changed and inflated the AOD values. In addition, with high humidity, particulates will likely be recorded as PM10 instead of the finer particulate fraction.

Figure 21. Correlation matrix of significant model inputs in Cleveland. The lower panel – scatterplots with loess smoothing; diagonal - histogram of data; upper panel – Pearson’s correlation coefficient. Plotted using pairs package in r (Becker, Chambers, & Wilks, 1988).

As expected there was a negative correlation between the PM$_{2.5}$ and AOD (7 observations) at RH less than 50% but a positive relationship (r = 0.40) existed when RH was high. It is expected that for six years, more than 7 days would have RH less than
50%, but this low count was a function of the availability of AOD and the subsequent date/time matching of observations.

For both cities, the height of the planetary boundary layer (MPBL) was significant for predicting PM$_{2.5}$. Lower MPBL is indicative of higher PM$_{2.5}$ and AOD (corrected) observations and vice versa. Interestingly, in Cincinnati, the raw or uncorrected AOD showed slight positive correlation with MPBL, indicating that the correction of the AOD values was important for this location. At both sites, there was improved correlation between AOD and PM$_{2.5}$ following correction. In Cincinnati, the correlation coefficient increased from 0.50 to 0.55 and while in Cleveland there was a 20% increase. These increases are in agreement with those done in Asia (Tsai et al., 2011) and Europe (Koelemeijer et al., 2006).

High wind speeds (ws) lead to dispersion and lower ambient aerosol concentration. This phenomena was revealed in the correlation plots shown above, but was only true when the AOD was corrected. Most studies examining the relationship between AOD and PM$_{2.5}$ found that air temperature was always significant, however this was not the case in Cincinnati. By including any of the temperature quantities (dew point, skin or air temperature) the testing and validation errors increased. Y. Liu et al. (2009) also found temperature to be non-robust variable but retained the variable in their analyses. Here it was not used.
Forecasting Model 1 – Multiple Linear Regression

The existing PM$_{2.5}$ forecasting model that was developed by the Center for Air Quality, includes all the variables listed above except for AOD in a multiple linear regression (MLR) model. The existing model also includes upper air temperature (700mb and 850mb), potential height and wild fire alert. That model operated on a six month and often underpredicts episodes. The average correlation coefficient between the predicted and observed PM$_{2.5}$ is 0.35. Using newly defined variable combinations, seasonal MLRs were created to first reduce the uncertainty produced by seasonality and secondly to explore the benefits of including AOD in forecasting models. For performance evaluation only the validation datasets (not used to train or test the model) were used.

Following the step-wise regression, AOD emerged as a significant positive predictor variable in all seasons. Equations 7 to 10 show the derived PM$_{2.5}$ prediction models in Cincinnati. The variance explained by the summer model was approximately 61% and represented the highest performance. In winter, however, the MLR for Cincinnati accounted for only 22% of the variance. This low model performance is a function of low data availability of data in the colder seasons.

**Summer:**

$$\text{PM}_{2.5} (\mu g / m^3) = -0.21 \times \text{RH} + 0.92 \times \text{VWND} - 17.38 \times \text{VIS} + 18.47 \times \text{AODc} + 52.73$$

(7)

**Fall:**

$$\text{PM}_{2.5} (\mu g / m^3) = -0.43 \times \text{VWND} - 7.98 \times \text{ws} - 10.56 \times \text{VIS} + 19.11 \times \text{AODc} + 32.23$$

(8)
\[ \text{Spring:} \]
\[ \text{PM}_{2.5}(\mu g / m^3) = -0.006 \times \text{HPBL} + 0.068 \times \text{WIND} - 7.98 \times \text{VIS} + 13.81 \times \text{AODc} + 30.68 \]  
(9)

\[ \text{Winter:} \]
\[ \text{PM}_{2.5}(\mu g / m^3) = 0.28 \times \text{RH} + 39.55 \times \text{AODc} - 9.79 \]  
(10)

Figure 22. Correlation matrix of significant model inputs in Cincinnati. The lower panel – scatterplots with loess smoothing; diagonal - histogram of data; upper panel – Pearson’s correlation coefficient. Plotted using pairs package in r (Becker et al., 1988).

Furthermore, besides AOD, only RH was significant for the model fit; this could mean that the covariates selected were not suitable for the winter season. Additional research exploring this point is therefore necessary. Figure 23 shows time-series and
scatterplots of the predicted and observed PM$_{2.5}$. It can be seen clearly that for all seasons the models are under predicting PM$_{2.5}$. The performance metric shown in Table 5 shows also confirms this.

![Image](image-url)

*Figure 23. Comparison of MLR-derived PM$_{2.5}$ with observed concentration at the Cincinnati study site. The blue line in the scatter plot represents the line of best fit and the vertical shows the standard deviation.*

<table>
<thead>
<tr>
<th>Season</th>
<th>N</th>
<th>FAC2</th>
<th>MB</th>
<th>MGE</th>
<th>NMB</th>
<th>NMGE</th>
<th>RMSE</th>
<th>r</th>
<th>COE</th>
<th>IOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
<td>140</td>
<td>0.92</td>
<td>0.20</td>
<td>4.76</td>
<td>0.01</td>
<td>0.30</td>
<td>5.99</td>
<td>0.56</td>
<td>0.14</td>
<td>0.57</td>
</tr>
<tr>
<td>Spring</td>
<td>122</td>
<td>0.93</td>
<td>0.33</td>
<td>4.69</td>
<td>0.02</td>
<td>0.28</td>
<td>5.75</td>
<td>0.66</td>
<td>0.17</td>
<td>0.59</td>
</tr>
<tr>
<td>Summer</td>
<td>194</td>
<td>0.99</td>
<td>-0.97</td>
<td>3.79</td>
<td>-0.05</td>
<td>0.20</td>
<td>4.96</td>
<td>0.78</td>
<td>0.37</td>
<td>0.68</td>
</tr>
<tr>
<td>Winter</td>
<td>54</td>
<td>0.93</td>
<td>0.70</td>
<td>4.84</td>
<td>0.05</td>
<td>0.32</td>
<td>5.76</td>
<td>0.47</td>
<td>0.03</td>
<td>0.51</td>
</tr>
</tbody>
</table>

A model of high performance has coefficient of efficiency (COE) values approaching unity (Legates & McCabe, 1999, 2013). The summer model as was
indicated by the r-squared value shows modest performance, but the COE value of 0.03 for the winter model implies that the model is no more capable of forecasting observed PM$_{2.5}$ than is the sample mean. Despite this, the fraction of predictions within a factor of two of observations (FAC2), which is a robust measure of performance due its resistance to outliers, indicated above average model performance for all the models, particularly for the summer. Similarly, the index of agreement (IOA) statistics showed that the predicted PM$_{2.5}$ values were in relatively good agreement with the observations. The overall mean (±SD) for the predicted and observed concentrations are 16.96 ± 5.14 µg/m$^3$ and 17.11 ± 7.57 µg/m$^3$, respectively. Statistically, there is no significant difference between these two measures.

Similarly at the Cleveland study location, the difference between the overall means (±SD) was not significant (observed - 17.40 ± 9.33 µg/m$^3$ and predicted - 17.34 ± 7.53 µg/m$^3$). The model performances, however, were lower than those identified in Cincinnati. The MLRs (Equations 11 to 14) all show relatively the same agreement between the predicted and observed PM$_{2.5}$, except in winter (Table 6). Here the winter COE is negative, which is an indication that the model is not effective at predicting the variation present in the observed data. The line of best in the Figure 24 clearly shows this behavior. On the other hand, both the predicted and observed values showed similar temporal trends and as can be seen in the Figure 24, their ranges were within 10 units unlike the Cincinnati models. Except for the winter models, the MLRs outperformed the existing Ohio University MLR.
Summer:
\[ PM_{2.5} (\mu g/m^3) = 0.56 \text{airTemp} - 0.006 \text{HPBL} + 1.51 \text{vwnd} - 7.68 \text{ws} + 15.3 \text{AOD} - 123.6 \]

(11)

Fall:
\[ PM_{2.5} (\mu g/m^3) = 33.45 \text{AOD} - 0.28 \text{RH} - 8.01 \text{ws} - 36.33 \text{Vis} + 95.7 \]

(12)

Spring:
\[ PM_{2.5} (\mu g/m^3) = 1.74 \text{airTemp} - 1.10 \text{dewpt} + 0.61 \text{vwnd} + 0.52 \text{uwnd} - 15.8 \text{PRCP} + 10.3 \text{AOD} - 13.2 \text{Vis} - 150.1 \]

(13)

Winter:
\[ PM_{2.5} (\mu g/m^3) = 0.012 \text{MPBL} - 21.2 \text{Vis} + 55.72 \]

(14)

Figure 24. Comparison of MLR-derived PM$_{2.5}$ with observed concentration at the Cleveland monitoring station. The blue line shows the line of best fit and 95% confidence interval.
MLR- AQI classification

AirNow (http://airnow.gov/) supplies daily color coded indicators (Table 3) specifying of air quality exposure risk levels across the United States. As shown in Table 3, these colors corresponds to specific AQI and PM$_{2.5}$ ranges. Figure 26 and 26 exhibit the predicted AQI categories derived from the forecasted PM$_{2.5}$ concentrations. In both cities, the models failed to appropriately specify the ‘unhealthy for sensitive groups’ category in fall and spring while summer has the only correctly forecasted unhealthy category. In fall and spring there are three cases where the unhealthy conditions were forecasted as moderate and one in spring where a moderate day was labeled as unhealthy. Note that fewer unhealthy days were observed in Cincinnati than in Cleveland.

Table 6

<table>
<thead>
<tr>
<th>Season</th>
<th>N</th>
<th>FAC2</th>
<th>MB</th>
<th>MGE</th>
<th>NMB</th>
<th>NMGE</th>
<th>RMSE</th>
<th>r</th>
<th>COE</th>
<th>IOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
<td>142</td>
<td>0.80</td>
<td>1.41</td>
<td>5.78</td>
<td>0.08</td>
<td>0.35</td>
<td>7.21</td>
<td>0.68</td>
<td>0.20</td>
<td>0.60</td>
</tr>
<tr>
<td>Spring</td>
<td>112</td>
<td>0.90</td>
<td>-0.71</td>
<td>5.69</td>
<td>-0.04</td>
<td>0.31</td>
<td>7.27</td>
<td>0.66</td>
<td>0.24</td>
<td>0.62</td>
</tr>
<tr>
<td>Summer</td>
<td>120</td>
<td>0.89</td>
<td>-0.69</td>
<td>5.35</td>
<td>-0.04</td>
<td>0.30</td>
<td>6.84</td>
<td>0.68</td>
<td>0.31</td>
<td>0.65</td>
</tr>
<tr>
<td>Winter</td>
<td>27</td>
<td>0.89</td>
<td>-2.36</td>
<td>5.90</td>
<td>-0.16</td>
<td>0.39</td>
<td>7.25</td>
<td>0.46</td>
<td>-0.19</td>
<td>0.40</td>
</tr>
</tbody>
</table>

The overall accuracies were affected by misclassification between moderate and good categories. That is, large portions of the observed good classifications (x-axis) were predicted as moderate and vice versa. Seasonal accuracies are displayed as percentages in each quadrant of Figures 25 and 26.
Figure 25. Graphical representation of the accuracy of the MLR models in estimating the Cleveland’s AQI categories.

Figure 26. Graphical representation of the accuracy of the MLR models in estimating Cincinnati’s AQI categories.
Forecasting Model 2 – PCA and MLR

It is obvious from the correlation plots (Figure 21 and 22) and MLRs shown above, that there are marginal co-linearity and nonlinear relationship between the input variables. Therefore, this phase of the analysis was done to reduce the number of potential inputs (Appendix B), identify and group variables with similar characteristics. To do this principal component analysis was done using two PCs selection criteria, B1Backward and Kaiser-Guttman. Application of the B1Backward PC selection criterion ($\lambda \geq 0.70$) resulted in 5 principal components for the two cities, but maintaining the traditional Kaiser-Guttman criterion ($\lambda \geq 1$) reduced the significant PCs to 3. The Kaiser-Guttman method is the traditional method used for PC selection, but the restriction of the Eigenvalues is arbitrary, therefore we explored the benefits of using the B1Backward method. The selected PCs explained more than 70% of the total variance with B1Backward approach, but only 65% from Kaiser-Guttman. Scree plots are shown in Figures B6 and B7 in Appendix B. Only those variables with significant loads were included in the calculating the weight of the selected PCs.

Cleveland - PCA

As is shown in Table 7, the selected PCs contained high loadings (variable coefficient) on each original input variable. There is a general trend in the PC structure; all temperature are loaded in PC1 and variables relating to atmospheric stability are combined to define the second PC. The structure for PC3 when $\lambda \geq 0.70$ is influenced by the presence of atmospheric moisture content, that is relatively humidity and positive
precipitation are positively loaded while visibility has a negative coefficient. However, when \( \lambda \geq 1 \), PC3 contains mixed effects; precipitation and relative humidity is combined with AOD. PC4 for B1Backward selection approach linked the wind component and PC5 fused the height of the planetary boundary layer and visibility to a negatively loaded AOD (\( \lambda \geq 0.70 \)).

**Cincinnati - PCA**

Similar to the structure outlined in Table 7 for Cleveland, PC1 for Cincinnati groups temperature inputs and PC2 describes stability through wind speed and mixing height (Table 8) when \( \lambda \geq 0.70 \), but the Kaiser-Guttman method suggested AOD as an influential component. PC4 can be described as the aerosol loading component since it contains a highly loaded AOD modifies by the mixing height.

**Regression**

The newly formed PCs were included as independent variable in a step-wise MLR and seasonal equations derived. Compared with the stand-alone MLR, only improvement to the predictive capacity was observed for the winter seasons. Equations 15 (Kaiser-Guttman) and 16 (B1Backward) represent the PM\(_{2.5}\) winter predictive equations derived for Cincinnati. Both models includes only the PC component with AOD, mixing height and wind speed. This implies that if an MLR is used to estimate winter PM\(_{2.5}\) in Cincinnati, covariates can be limited to wind properties, AOD observations and mixing
height. Notice also that east-west horizontal winds (UWND), which is an indication that advection and potential near surface build up particulates are influential.

\[ PM_{2.5} = 15.27 - 3.061 \,(PC2) + 3.41 \,(PC3) \]  \hspace{1cm} (15)

\[ PM_{2.5} = 15.64 - 1.75 \,(PC2) + 5.92 \,(PC4) \]  \hspace{1cm} (16)

Table 7

*Selected PC for Cleveland: BIBackward (\(\lambda \geq 0.70\)) and Kaiser-Guttman (\(\lambda \geq 1\))*

<table>
<thead>
<tr>
<th></th>
<th>(\lambda \geq 0.70)</th>
<th>(\lambda \geq 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>0.995 airTemp + 0.972 DewPt + 0.939 TempSurf</td>
<td>0.989 airTemp + 0.98 DewPt + 0.988 TempSurf</td>
</tr>
<tr>
<td>PC2</td>
<td>0.886 wd + 0.338 UWND + 0.670 MPBL + 0.756 VWND</td>
<td>0.666 ws + 0.706 UWND + 0.540 wd + 0.740 VWND</td>
</tr>
<tr>
<td>PC3</td>
<td>0.806 RH – 0.743 VIS + 0.560 PRCP</td>
<td>0.70 RH – 0.508 AODc + 0.648 PRCP</td>
</tr>
<tr>
<td>PC4</td>
<td>0.901 wd + 0.747 UWND</td>
<td></td>
</tr>
<tr>
<td>PC5</td>
<td>0.503 MPBL + 0.326 VIS – 0.876 AOD</td>
<td></td>
</tr>
</tbody>
</table>
Table 8

**Selected PC for Cincinnati: B1Backward (λ ≥ 0.70) and Kaiser-Guttman (λ ≥ 1)**

<table>
<thead>
<tr>
<th></th>
<th>λ ≥ 0.70</th>
<th>λ ≥ 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>0.995 airTemp + 0.970 DewPt + 0.950TempSurf</td>
<td>0.995 airTemp + 0.970 DewPt + 0.959TempSurf</td>
</tr>
<tr>
<td>PC2</td>
<td>0.891 ws + 0.803MPBL</td>
<td>0521 wd + 0.686 UWND + 0.827MPBL − 0.646 AODc</td>
</tr>
<tr>
<td>PC3</td>
<td>0.838 RH + 744 PRCP</td>
<td>0.865RH + 0.662PRCP + 0.328ws</td>
</tr>
<tr>
<td>PC4</td>
<td>0.913 AODc − 0.368 HPBL</td>
<td></td>
</tr>
<tr>
<td>PC5</td>
<td>0.966 wd</td>
<td></td>
</tr>
</tbody>
</table>

The lowest root mean squared error (RSME) shown in Table 9, is also attributed to the winter models. The marginal increase of the COE statistic and correlation coefficient shown for the B1Backward model suggest better performance than the stand-alone MLR. In fact, the B1Backward models were better able to predict PM$_{2.5}$ than those developed using the Kaiser - Guttman PC selection approach. This is illustrated in Figures 27 and 28.

Neither of the models were able to accurately capture high levels of PM$_{2.5}$. The unhealthy category was consistently misrepresented and or not calculated at all (Figure 28). The fact that winter had no unhealthy days in either city bolsters the performances. In total there were 12 unhealthy days observed in Cincinnati, 6 of which were included in the validation dataset, 4 in training and 2 in testing. Similar proportional breakdown was used for the Cleveland dataset except each partition contains twice the observations.
Therefore, given the small number of unhealthy dat's in the data, it is not surprising that
the models did not perform well for the unhealthy category.

Table 9

*Seasonal model performance metrics for the Cincinnati*

<table>
<thead>
<tr>
<th>Season</th>
<th>N</th>
<th>FAC2</th>
<th>MB</th>
<th>MGE</th>
<th>NMB</th>
<th>NMGE</th>
<th>RMSE</th>
<th>r</th>
<th>COE</th>
<th>IOA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaiser</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fall</td>
<td>140</td>
<td>0.94</td>
<td>-0.08</td>
<td>5.00</td>
<td>-0.01</td>
<td>0.31</td>
<td>6.33</td>
<td>0.48</td>
<td>0.10</td>
<td>0.55</td>
</tr>
<tr>
<td>Spring</td>
<td>122</td>
<td>0.91</td>
<td>0.17</td>
<td>4.75</td>
<td>0.01</td>
<td>0.29</td>
<td>6.30</td>
<td>0.56</td>
<td>0.16</td>
<td>0.58</td>
</tr>
<tr>
<td>Summer</td>
<td>194</td>
<td>0.96</td>
<td>-0.77</td>
<td>4.88</td>
<td>-0.04</td>
<td>0.26</td>
<td>6.56</td>
<td>0.55</td>
<td>0.18</td>
<td>0.59</td>
</tr>
<tr>
<td>Winter</td>
<td>54</td>
<td>0.85</td>
<td>1.58</td>
<td>5.07</td>
<td>0.10</td>
<td>0.34</td>
<td>6.07</td>
<td>0.43</td>
<td>-0.02</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1Backward</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fall</td>
<td>140</td>
<td>0.93</td>
<td>-0.33</td>
<td>4.55</td>
<td>-0.02</td>
<td>0.29</td>
<td>5.96</td>
<td>0.56</td>
<td>0.18</td>
<td>0.59</td>
</tr>
<tr>
<td>Spring</td>
<td>122</td>
<td>0.93</td>
<td>0.21</td>
<td>4.72</td>
<td>0.01</td>
<td>0.28</td>
<td>5.90</td>
<td>0.63</td>
<td>0.17</td>
<td>0.58</td>
</tr>
<tr>
<td>Summer</td>
<td>194</td>
<td>0.95</td>
<td>-0.90</td>
<td>4.60</td>
<td>-0.05</td>
<td>0.24</td>
<td>6.11</td>
<td>0.63</td>
<td>0.23</td>
<td>0.62</td>
</tr>
<tr>
<td>Winter</td>
<td>54</td>
<td>0.96</td>
<td>0.72</td>
<td>4.54</td>
<td>0.05</td>
<td>0.30</td>
<td>5.36</td>
<td>0.54</td>
<td>0.09</td>
<td>0.54</td>
</tr>
</tbody>
</table>
Figure 27. Kaiser-Guttman: comparison of derived PM$_{2.5}$ with observation made at the Cincinnati monitoring station (top). Graphical representation of the accuracy of the 3 component models in estimating Cincinnati’s AQI categories (bottom).
To explore the performance of the models for these high levels of PM$_{2.5}$ the training sample size will have to be increased. This could be done by increasing the historical datasets used. In Cleveland AOD was not a significant component in any of the
winter models. Time-series and model performance estimates for the other seasons are shown in Appendix B.

*Forecasting Model 3 - Neural Network*

The final models evaluated in the study are artificial neural networks. The general schematic for each of the models developed is shown in Figure 20. As was the case for the MLR, and PCA, models for individual seasons are examined. So far the MLR and PCA models, while being able to modestly predictive PM$_{2.5}$, consistently misclassified peak concentrations. This was partly due to the structure and combination of input variables in each model; hence, the neural network was proposed as a way to breakdown this complexity. Emphasis was placed on the model’s capability to estimate higher air quality risk levels even though false alarm rates and success rates of the lower level air quality index are important.

The contingency plots shown in Figure 30, show that for both cities, the models were able to detect and properly identify unhealthy days. Of the 5 unhealthy days selected in the Cincinnati validation datasets, 3 were accurately placed. Similarly, in Cleveland the unhealthy categories was 45% accurate. These were significant improvements over the previously discussed MLR and PCA models. An overall accuracy of 77% was obtained in Cincinnati. Cleveland was marginally less with 74%, but showed improvements in fall and summer’s “good” and “moderate” categories, respectively. Contrarily, the statistics (Table 10) show that the accuracies for winter season did not improve.
Figure 29. 24-hour estimation of PM$_{2.5}$ modeled using neural networks. Cleveland (top row) and Cincinnati (bottom row).
For all the network models (Appendix C), AOD was an important predictor. The thickness of the connections suggested the importance of each variable. During, fall and summer seasons in Cleveland, the height of mixing layer was most important while in spring AOD was the least important significant variable. The result of this can be seen in Figure 29 where spring had the lowest accuracy in the good and moderate categories. For the fall season in Cincinnati, almost 50% of the moderate category were modeled as good. Like the other models, winter had the highest accuracy, 80%.

Summary

In total 8 separate models (2 per city, one per season) were developed. The performances varied by city, season and model type. Therefore, in order to assess the overall accuracies, Taylor diagrams (Taylor, 2001), which concisely compared correlation, centered RSME and variances were drawn. In Cleveland (Figure 31), the standalone MLR was the best model for spring and winter. It captured the variability in the observations, which is shown on the diagram as the point that falls the closest to the observed variability arc (starting from the purple point labeled observed). The five component PCA had a higher correlation and lower RSME in winter, but the difference was not sufficient to make it better than the MLR. In summer and fall (autumn), the neural network model explained the most of the observed variance, has the highest correlation coefficient and lowest error. The PCAs selected using tradition Kaiser-Guttman method was least successful model.
In Cincinnati, the neural network and MLR models were the most useful for the summer and fall seasons. In spring both models behaved nearly the same and in winter the PCA 5 model had an edge over the other models due to higher correlation and lower error. The differences among the winter models are negligible.

Table 10

*Neural network performance statistics divided into four seasons for each city*

<table>
<thead>
<tr>
<th>Season</th>
<th>N</th>
<th>FAC2</th>
<th>MB</th>
<th>MGE</th>
<th>NMB</th>
<th>NMGE</th>
<th>RMSE</th>
<th>r</th>
<th>COE</th>
<th>IOA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cleveland</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fall</td>
<td>142</td>
<td>0.82</td>
<td>3.19</td>
<td>5.63</td>
<td>0.19</td>
<td>0.34</td>
<td>6.94</td>
<td>0.78</td>
<td>0.22</td>
<td>0.61</td>
</tr>
<tr>
<td>Spring</td>
<td>112</td>
<td>0.82</td>
<td>-0.14</td>
<td>6.12</td>
<td>-0.01</td>
<td>0.34</td>
<td>7.84</td>
<td>0.58</td>
<td>0.18</td>
<td>0.59</td>
</tr>
<tr>
<td>Summer</td>
<td>120</td>
<td>0.92</td>
<td>-0.37</td>
<td>5.09</td>
<td>-0.02</td>
<td>0.28</td>
<td>6.91</td>
<td>0.68</td>
<td>0.34</td>
<td>0.67</td>
</tr>
<tr>
<td>Winter</td>
<td>27</td>
<td>0.78</td>
<td>-2.84</td>
<td>5.54</td>
<td>-0.19</td>
<td>0.37</td>
<td>7.15</td>
<td>0.29</td>
<td>-0.12</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cincinnati</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fall</td>
<td>136</td>
<td>0.87</td>
<td>0.54</td>
<td>4.87</td>
<td>0.03</td>
<td>0.30</td>
<td>6.05</td>
<td>0.57</td>
<td>0.16</td>
<td>0.58</td>
</tr>
<tr>
<td>Spring</td>
<td>119</td>
<td>0.87</td>
<td>1.48</td>
<td>5.00</td>
<td>0.09</td>
<td>0.30</td>
<td>6.29</td>
<td>0.57</td>
<td>0.16</td>
<td>0.58</td>
</tr>
<tr>
<td>Summer</td>
<td>181</td>
<td>0.96</td>
<td>-0.41</td>
<td>4.04</td>
<td>-0.02</td>
<td>0.22</td>
<td>5.08</td>
<td>0.77</td>
<td>0.36</td>
<td>0.68</td>
</tr>
<tr>
<td>Winter</td>
<td>53</td>
<td>0.94</td>
<td>-1.42</td>
<td>5.34</td>
<td>-0.09</td>
<td>0.34</td>
<td>6.83</td>
<td>0.42</td>
<td>0.10</td>
<td>0.55</td>
</tr>
</tbody>
</table>
Figure 30. Contingency plot of predicted 24-hour estimation of PM$_{2.5}$ (neural network). Cleveland (top) and Cincinnati (bottom).
CONCLUSION

This work aimed to show that aerosol optical depth included in forecasting models is viable predictor variable for PM$_{2.5}$ as measured by standard ground-based monitors in the region. Three types of forecasting models, MLR, PCA/MLR and neural network, were developed, trained, tested and validated using historical and forecasted meteorological parameters and aerosol loadings. The models were developed specifically for Cincinnati and Cleveland metropolitan areas after trend analyses revealed consistently high PM$_{2.5}$ concentration levels at both locations.

Figure 31. Taylor diagram displaying statistical comparison the models developed for Cleveland
A look at ORV region showed that AOD and PM$_{2.5}$ were in relatively good agreement both spatial and temporally. Consequently, to narrow the scope of the research only monitoring stations in Ohio were studied. Ohio was also considered to be representative of the region based on the diversity of the five EPA defined monitoring sectors. There are regions bordering the Great Lakes, juxtaposing the Ohio River and those in the center with typical PM$_{2.5}$ sources. In addition, other studies have shown that the state is both a source and receptor location for PM$_{2.5}$.

Figure 32. Taylor diagram displaying statistical comparison the models developed for Cincinnati
Trend analysis identified general decline in average PM$_{2.5}$ levels across the state, but only in specific areas. On average, there was a net decline of $\sim$2µg/m$^3$ of fine particulate concentration over the six year study period. The sharpest decline occurred in Franklin County. In the Northeast, specifically Cuyahoga county, annual averages were greater than 15µg/m$^3$ and showed no significant change over time. Sparse distribution of ambient monitoring stations makes air quality assessment difficult and such forecasting models have been employed by states and the EPA as a supporting tool.

The use of remote sensing to support ground-based measurement is an ongoing research. Satellite data offers a way to support air quality management in the state through frequent synoptic observation of atmospheric aerosol loadings. As shown here, PM$_{2.5}$-AOD correlations are low when one to one comparisons are done, but following modifications by local meteorological parameters, AOD can be an effective tool for PM$_{2.5}$ estimations. In Cincinnati, it was shown that neural networks and MLR worked best for summer and fall season, but of the 4 models, none was able to fully describe the observations in winter. The same pattern was observed in Cleveland, but the PCA structure breakdown revealed that AOD combined with local wind properties and mixing height improved model accuracies for winter.

Further explorations of potential model input parameters are required to enhance model performances, especially in the cold season. Since observations are lowest in the winter for AOD, a wider range of historical data may be required in the future. In addition, inconsistent satellite measurements limits the short term use of AOD, but as was done for this study, combining collocated complementary satellite data, increased the
power of analysis. The next step in this research area would be to determine if other explanatory variables affect the AOD-PM$_{2.5}$ relationship. The new push by the EPA to incorporate more remote-sensing products into forecasting models will add more focus to the area and answer other questions regarding the complex AOD-PM$_{2.5}$ relationship.
REFERENCES


...


White, E.M. (2009). *Source attribution, physicochemical properties and spatial distribution of wet deposited mercury to the Ohio river valley.* (PhD), University of Michigan.


APPENDIX A: REGIONAL HOURLY PM2.5

Boxplots showing PM$_{2.5}$ distribution at monitoring sites throughout the state of Ohio.

*Figure A1.* Boxplots of hourly PM$_{2.5}$ concentration at monitoring sites in central Ohio.
Figure A2. Boxplots of hourly PM$_{2.5}$ concentration at monitoring sites in SE Ohio

Figure A3. Boxplots of hourly PM$_{2.5}$ concentration at monitoring sites in NW Ohio
Figure A4. PM$_{2.5}$ 24-hour 98th percentile, averaged over 3 years northeast Ohio

Figure A5. PM$_{2.5}$ 24-hour average (98$^{th}$ percentile) in southeast Ohio
Figure A6. PM25 24-hour average (98th percentile) in central Ohio

Figure A7. PM25 24-hour average (98th percentile) in northwest Ohio
Figure A8. PM25 24-hour average (98th percentile) in southwestern Ohio
APPENDIX B: POTENTIAL MODEL INPUT VARIABLES

Table B1. Descriptive statistics of potential variables separated into categories

<table>
<thead>
<tr>
<th></th>
<th>Temperature</th>
<th>Surface Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>airTemp</td>
<td>SurfTemp</td>
</tr>
<tr>
<td>Mean</td>
<td>289.2</td>
<td>289.8</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.269</td>
<td>0.277</td>
</tr>
<tr>
<td>Median</td>
<td>291.5</td>
<td>291.6</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>9.361</td>
<td>9.648</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0.677</td>
<td>-0.762</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.529</td>
<td>-0.424</td>
</tr>
<tr>
<td>Minimum</td>
<td>260.8</td>
<td>262.4</td>
</tr>
<tr>
<td>Maximum</td>
<td>305.5</td>
<td>309.1</td>
</tr>
<tr>
<td>Confidence Level (95.0%)</td>
<td>0.528</td>
<td>0.544</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Stability</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>wd</td>
<td>ws</td>
</tr>
<tr>
<td>Mean</td>
<td>194.1</td>
<td>0.581</td>
</tr>
<tr>
<td>Standard Error</td>
<td>2.882</td>
<td>0.007</td>
</tr>
<tr>
<td>Median</td>
<td>210.0</td>
<td>0.509</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>100.3</td>
<td>0.258</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0.926</td>
<td>2.290</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.369</td>
<td>1.258</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.000</td>
<td>0.149</td>
</tr>
<tr>
<td>Maximum</td>
<td>360.0</td>
<td>2.006</td>
</tr>
<tr>
<td>Confidence Level (95.0%)</td>
<td>5.655</td>
<td>0.015</td>
</tr>
</tbody>
</table>
Figure B1. Correlogram of potential model inputs- Cincinnati
Figure B2. Correlogram of potential model inputs- Cleveland
Model Performance – PCA and MLR in Cleveland

Figure B3. Time series plots of observed PM$_{2.5}$ and predicted PM2.5 – 3 PCA in Cleveland

Figure B4. Time series plots of observed PM$_{2.5}$ and predicted PM2.5 – 5 PCA in Cleveland, vertical bars represent ±1 standard deviation.
Figure B5. PM$_{2.5}$ trends conditioned by wind direction at the Cleveland site.
Figure B6. Scree plot for Cincinnati PCA selection ($\lambda=1$)

Figure B7. Scree plot for Cleveland PCA selection
APPENDIX C: NEURAL NETWORKS AND CODES

Cincinnati

Figure C1. Neural network used for the Cincinnati spring season

Neural Network code –spring

```xml
<?xml version="1.0" encoding="UTF-8"?
<PMML version="4.1"
xmlns="http://www.dmg.org/PMML-4_1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema">
<Header copyright="(C) Copyright IBM Corp. 1994, 2013">
  <Application name="IBM SPSS Modeler Common" version="16.0.0.0"/>
  <Timestamp>Tue Jul 08 22:22:42 2014</Timestamp>
</Header>
<MiningBuildTask>
  <Extension extender="spss.com" name="split-info">
    <DataField dataType="string" name="Season" otype="categorical">
      <Value value="Spring"/>
    </DataField>
  </Extension>
</MiningBuildTask>
<DataDictionary numberOfFields="8">
  <DataField dataType="double" displayName="AODc" name="AODc" otype="continuous"/>
```
```
<DataField dataType="double" displayName="HPBL" name="HPBL" optype="continuous"/>
<DataField dataType="double" displayName="PM25" name="PM25" optype="continuous"/>
<DataField dataType="double" displayName="PRCP" name="PRCP" optype="continuous"/>
<DataField dataType="double" displayName="VIS" name="VIS" optype="continuous"/>
<DataField dataType="double" displayName="uwnd" name="uwnd" optype="continuous"/>
<DataField dataType="double" displayName="vwnd" name="vwnd" optype="continuous"/>
<DataField dataType="double" displayName="ws" name="ws" optype="continuous"/>
</DataDictionary><TransformationDictionary>
<DerivedField dataType="double" name="AODcNorm" optype="continuous">
<NormContinuous field="AODc">
<LinearNorm norm="-1.17072401948097" orig="0.012879359516492"/>
<LinearNorm norm="3.34719266980786" orig="0.657671455344572"/>
</NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="HPBLNorm" optype="continuous">
<NormContinuous field="HPBL">
<LinearNorm norm="-1.37348117933508" orig="455.1"/>
<LinearNorm norm="3.96158870776459" orig="1852.8"/>
</NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="PM25Norm" optype="continuous">
<NormContinuous field="PM25">
<LinearNorm norm="-1.83031249053993" orig="4.566667"/>
<LinearNorm norm="2.96984140227963" orig="37.756"/>
</NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="PRCPNorm" optype="continuous">
<NormContinuous field="PRCP">
<LinearNorm norm="-0.396366886979895" orig="0"/>
<LinearNorm norm="5.3568866102612" orig="2.02"/>
<TransformationDictionary>
<DerivedField dataType="double" name="VISNorm" optype="continuous">
  <NormContinuous field="VIS">
    <LinearNorm norm="-2.95067116212212" orig="0.7889"/>
    <LinearNorm norm="0.824918819518013" orig="1.61"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="uwndNorm" optype="continuous">
  <NormContinuous field="uwnd">
    <LinearNorm norm="-2.01535440629688" orig="-4.748"/>
    <LinearNorm norm="2.28674066485568" orig="6.135"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="vwndNorm" optype="continuous">
  <NormContinuous field="vwnd">
    <LinearNorm norm="-2.81662853586352" orig="-7.242"/>
    <LinearNorm norm="2.50167453805924" orig="6.788"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="wsNorm" optype="continuous">
  <NormContinuous field="ws">
    <LinearNorm norm="-1.45485218911295" orig="0.2623644"/>
    <LinearNorm norm="3.58032893895051" orig="1.440432"/>
  </NormContinuous>
</DerivedField>

</TransformationDictionary>
<NeuralNetwork activationFunction="tanh" algorithmName="MLP" functionName="regression">
  <Extension extender="spss.com" name="modellID" value="0"/>
  <MiningSchema>
    <MiningField importance="0.212367276247078" name="AODc"/>
    <MiningField importance="0.161709357244376" name="HPBL"/>
    <MiningField importance="0.0904783211177229" name="PRCP"/>
    <MiningField importance="0.0805512587817674" name="VIS"/>
  </MiningSchema>
</NeuralNetwork>
<MiningField importance="0.0294996478264632" name="uwnd"/>
<MiningField importance="0.305743782753827" name="vwnd"/>
<MiningField importance="0.119650356028766" name="ws"/>
<MiningField name="PM25" usageType="predicted"/>

<FieldRef field="PRCPNorm"/>
</DerivedField>
</NeuralInput>
<NeuralInput id="3">

<DerivedField data_type="double" otype="continuous">
<FieldRef field="VISNorm"/>
</DerivedField>
</NeuralInput>
<NeuralInput id="4">

<DerivedField data_type="double" otype="continuous">
<FieldRef field="uwndNorm"/>
</DerivedField>
</NeuralInput>
<NeuralInput id="5">

<DerivedField data_type="double" otype="continuous">
<FieldRef field="vwndNorm"/>
</DerivedField>
</NeuralInput>
<NeuralInput id="6">

<DerivedField data_type="double" otype="continuous">
<FieldRef field="wsNorm"/>
</DerivedField>
</NeuralInput>
</NeuralInputs><NeuralLayer number_of_neurons="3">

<Neuron bias="0.543903458327636" id="7">
<Con from="0" weight="-0.114203146027586"/>
<Con from="1" weight="-0.424093546888462"/>
<Con from="2" weight="0.292273771322259"/>
<Con from="3" weight="-0.539183810797606"/>
<Con from="4" weight="-0.113654001439223"/>
<Con from="5" weight="0.231754086844965"/>
<Con from="6" weight="0.183572509678078"/>
</Neuron>

<Neuron bias="-0.200377729056161" id="8">
<Con from="0" weight="-0.450784673735331"/>
<Con from="1" weight="-0.145296739549766"/>
<Con from="2" weight="0.11504458256537"/>
<Con from="3" weight="0.188024967138575"/>
<Con from="4" weight="-0.0774534529050804"/>
<Con from="5" weight="-0.433574395306672"/>
<Con from="6" weight="-0.109353012379399"/>
</Neuron>

<Neuron bias="-0.260520543206239" id="9">
<Con from="0" weight="0.432528502410113"/>
</Neuron>
<Con from="1" weight="-0.290347597846449"/>
<Con from="2" weight="-0.233992771971513"/>
<Con from="3" weight="0.023070752890678"/>
<Con from="4" weight="0.097419105486072"/>
<Con from="5" weight="0.388843058062639"/>
<Con from="6" weight="-0.530595533295128"/>
</Neuron>
</NeuralLayer><NeuralLayer activationFunction="identity" numberOfNeurons="1">
    <Neuron bias="-0.238586414646495" id="10">
        <Con from="7" weight="0.332197938739355"/>
        <Con from="8" weight="-0.41510692279286"/>
        <Con from="9" weight="0.729147401385034"/>
    </Neuron>
</NeuralLayer>
</NeuralOutputs>
</NeuralNetwork></PMML>
**Figure C2.** Neural network used for Cincinnati summer season

**Neural Network code – summer**

```xml
<?xml version="1.0" encoding="UTF-8"?>
<PMML version="4.1"
xmlns="http://www.dmg.org/PMML-4_1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema">
  <Header copyright="(C) Copyright
IBM Corp. 1994, 2013">
    <Application name="IBM SPSS Modeler Common" version="16.0.0.0"/>
    <Timestamp>Tue Jul 08 22:22:42 2014</Timestamp>
  </Header>
  <MiningBuildTask>
    <Extension extender="spss.com" name="split-info">
      <DataField dataType="string" name="Season" optype="categorical">
        <Value value="Summer"/>
      </DataField>
    </Extension>
  </MiningBuildTask>
  <DataDictionary numberOfFields="8">
    <DataField dataType="double" displayName="AODc" name="AODc" optype="continuous"/>
    <DataField dataType="double" displayName="HPBL" name="HPBL" optype="continuous"/>
    <DataField dataType="double" displayName="PM25" name="PM25" optype="continuous"/>
  </DataDictionary>
</PMML>
```
<DataField dataType="double" displayName="PRCP" name="PRCP" optype="continuous"/>
<DataField dataType="double" displayName="VIS" name="VIS" optype="continuous"/>
<DataField dataType="double" displayName="uwnd" name="uwnd" optype="continuous"/>
<DataField dataType="double" displayName="vwnd" name="vwnd" optype="continuous"/>
<DataField dataType="double" displayName="ws" name="ws" optype="continuous"/>
</DataDictionary><TransformationDictionary>
<DerivedField dataType="double" name="AODcNorm" optype="continuous">
  <NormContinuous field="AODc">
    <LinearNorm norm="-1.4324420695782" orig="-0.0253677294236313"/>
    <LinearNorm norm="2.29132972161089" orig="0.406734422082775"/>
  </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="HPBLNorm" optype="continuous">
  <NormContinuous field="HPBL">
    <LinearNorm norm="-1.80871600955038" orig="528.6"/>
    <LinearNorm norm="3.05262566894193" orig="1616.7"/>
  </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="PM25Norm" optype="continuous">
  <NormContinuous field="PM25">
    <LinearNorm norm="-1.73746737603765" orig="4.175"/>
    <LinearNorm norm="3.47564851785572" orig="44.891667"/>
  </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="PRCPNorm" optype="continuous">
  <NormContinuous field="PRCP">
    <LinearNorm norm="-0.29468736034213" orig="0"/>
    <LinearNorm norm="7.09123860482287" orig="1.89"/>
  </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="VISNorm" optype="continuous">
  
</DerivedField>
<NormContinuous field="VIS">
  <LinearNorm norm="-2.25192616254399" orig="0.8694"/>
  <LinearNorm norm="1.04086759053686" orig="1.61"/>
</NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="uwndNorm" optype="continuous">
  <NormContinuous field="uwnd">
    <LinearNorm norm="-2.3946838562353" orig="-3.729"/>
    <LinearNorm norm="2.1251962496116" orig="5.327"/>
  </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="vwndNorm" optype="continuous">
  <NormContinuous field="vwnd">
    <LinearNorm norm="-2.02730086208212" orig="-3.577"/>
    <LinearNorm norm="2.01808861038009" orig="4.883"/>
  </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="wsNorm" optype="continuous">
  <NormContinuous field="ws">
    <LinearNorm norm="-1.21658944233575" orig="0.2623644"/>
    <LinearNorm norm="4.4917378337767" orig="1.80054"/>
  </NormContinuous>
</DerivedField>
</TransformationDictionary>
<NeuralNetwork activationFunction="tanh" algorithmName="MLP" functionName="regression"><Extension extender="spss.com" name="modelID" value="1"/></NeuralNetwork>

<MiningSchema>
  <MiningField importance="0.216963940151378" name="AODc"/>
  <MiningField importance="0.107167591278767" name="HPBL"/>
  <MiningField importance="0.150580262346014" name="PRCP"/>
  <MiningField importance="0.18519728263956" name="VIS"/>
  <MiningField importance="0.133636376881604" name="uwnd"/>
  <MiningField importance="0.185928794786937" name="vwnd"/>
  <MiningField importance="0.0205257522913435" name="ws"/>
  <MiningField name="PM25" usageType="predicted"/>
</MiningSchema>
</ModelExplanation>
</PredictiveModelQuality dataUsage="training" meanSquaredError="24.8209324693379" r-
<FieldRef field="PM25Norm"/>
</DerivedField>
</NeuralOutput>
</NeuralOutputs></NeuralNetwork></PMML>
Neural Network code – winter

<?xml version="1.0" encoding="UTF-8"?>
<PMML version="4.1"
xmlns="http://www.dmg.org/PMML-4_1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema">
<Header copyright="(C) Copyright IBM Corp. 1994, 2013">
  <Application name="IBM SPSS Modeler Common" version="16.0.0.0"/>
  <Timestamp>Tue Jul 08 22:22:42 2014</Timestamp>
</Header>
<MiningBuildTask>
  <Extension extender="spss.com" name="split-info">
    <DataField dataType="string" name="Season" optype="categorical">
      <Value value="Winter"/>
    </DataField>
  </Extension>
</MiningBuildTask>
<DataDictionary numberOfFields="8">
  <DataField dataType="double" displayName="AODc" name="AODc" optype="continuous"/>
  <DataField dataType="double" displayName="HPBL" name="HPBL" optype="continuous"/>
  <DataField dataType="double" displayName="PM25" name="PM25" optype="continuous"/>

Figure C3. Neural network used for Cincinnati winter season
<DataDictionary>
  <DataField dataType="double" displayName="PRCP" name="PRCP" optype="continuous"/>
  <DataField dataType="double" displayName="VIS" name="VIS" optype="continuous"/>
  <DataField dataType="double" displayName="uwnd" name="uwnd" optype="continuous"/>
  <DataField dataType="double" displayName="vwnd" name="vwnd" optype="continuous"/>
  <DataField dataType="double" displayName="ws" name="ws" optype="continuous"/>
</DataDictionary>
<TransformationDictionary>
  <DerivedField dataType="double" name="AODcNorm" optype="continuous">
    <NormContinuous field="AODc">
      <LinearNorm norm="-0.831282358000236" orig="0.0127796652763123"/>
      <LinearNorm norm="4.48519949182338" orig="0.721118652463549"/>
    </NormContinuous>
  </DerivedField>
  <DerivedField dataType="double" name="HPBLNorm" optype="continuous">
    <NormContinuous field="HPBL">
      <LinearNorm norm="-1.47123931650222" orig="224.2"/>
      <LinearNorm norm="2.26681229954283" orig="1841.75"/>
    </NormContinuous>
  </DerivedField>
  <DerivedField dataType="double" name="PM25Norm" optype="continuous">
    <NormContinuous field="PM25">
      <LinearNorm norm="-1.33387708586121" orig="4.004167"/>
      <LinearNorm norm="4.11953604581449" orig="49.325"/>
    </NormContinuous>
  </DerivedField>
  <DerivedField dataType="double" name="PRCPNorm" optype="continuous">
    <NormContinuous field="PRCP">
      <LinearNorm norm="-0.377141310231389" orig="0"/>
      <LinearNorm norm="4.74037616090838" orig="0.21"/>
    </NormContinuous>
  </DerivedField>
</TransformationDictionary>
<DerivedField dataType="double" name="VISNorm" optype="continuous">
    <NormContinuous field="VIS">
        <LinearNorm norm="-3.41303272038314" orig="0.8533"/>
        <LinearNorm norm="0.566486824824392" orig="1.61"/>
    </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="uwndNorm" optype="continuous">
    <NormContinuous field="uwnd">
        <LinearNorm norm="-1.4781909800983" orig="-2.432"/>
        <LinearNorm norm="2.44141065008218" orig="9.915"/>
    </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="vwndNorm" optype="continuous">
    <NormContinuous field="vwnd">
        <LinearNorm norm="-2.31661009241717" orig="-4.104"/>
        <LinearNorm norm="1.97991317889193" orig="5.818"/>
    </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="wsNorm" optype="continuous">
    <NormContinuous field="ws">
        <LinearNorm norm="-1.60324926807048" orig="0.1491876"/>
        <LinearNorm norm="2.24226947396967" orig="1.388988"/>
    </NormContinuous>
</DerivedField>

</TransformationDictionary><NeuralNetwork activationFunction="tanh" algorithmName="MLP" functionName="regression"><Extension extender="spss.com" name="modelID" value="2"/></NeuralNetwork>

<MiningSchema>
    <MiningField importance="0.173304317569465" name="AODc"/>
    <MiningField importance="0.288159639289459" name="HPBL"/>
    <MiningField importance="0.0748383162366335" name="PRCP"/>
    <MiningField importance="0.0518965531349453" name="VIS"/>
    <MiningField importance="0.101994900213109" name="uwnd"/>
    <MiningField importance="0.179337916850403" name="vwnd"/>
    <MiningField importance="0.130468356705985" name="ws"/>
<NeuralInput id="4">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="uwndNorm"/>
  </DerivedField>
</NeuralInput>

<NeuralInput id="5">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="vwndNorm"/>
  </DerivedField>
</NeuralInput>

<NeuralInput id="6">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="wsNorm"/>
  </DerivedField>
</NeuralInput>
</NeuralInputs>

<NeuralLayer numberOfNeurons="3">
  <Neuron bias="0.5346556468686677" id="7">
    <Con from="0" weight="-0.392539868364111"/>
    <Con from="1" weight="0.3399115777577448"/>
    <Con from="2" weight="-0.468908843584359"/>
    <Con from="3" weight="0.213719077641144"/>
    <Con from="4" weight="-0.17054098425433"/>
    <Con from="5" weight="0.512809299631044"/>
    <Con from="6" weight="0.0345365682151169"/>
  </Neuron>
  <Neuron bias="0.0208947693463415" id="8">
    <Con from="0" weight="-0.246454724809155"/>
    <Con from="1" weight="-0.0404620121698827"/>
    <Con from="2" weight="0.292123397346586"/>
    <Con from="3" weight="0.118247155332938"/>
    <Con from="4" weight="-0.591069290880114"/>
    <Con from="5" weight="-0.0996212060563266"/>
    <Con from="6" weight="-0.228254155954346"/>
  </Neuron>
  <Neuron bias="0.425163361476734" id="9">
    <Con from="0" weight="0.0609626192599535"/>
    <Con from="1" weight="0.647255151765421"/>
    <Con from="2" weight="-0.0515028925146908"/>
    <Con from="3" weight="-0.0842406617011875"/>
    <Con from="4" weight="0.460400523617864"/>
    <Con from="5" weight="-0.369121431140229"/>
    <Con from="6" weight="0.484650831669569"/>
  </Neuron>
</NeuralLayer>
Figure C4. Neural network used for Cincinnati fall season

**Neural Network code – fall**

```xml
<?xml version="1.0" encoding="UTF-8"?><PMML version="4.1"
xmlns="http://www.dmg.org/PMML-4_1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema"><Header copyright="(C) Copyright IBM Corp. 1994, 2013">
  <Application name="IBM SPSS Modeler Common" version="16.0.0.0"/>
  <Timestamp>Tue Jul 08 22:22:42 2014</Timestamp>
</Header><MiningBuildTask><Extension extender="spss.com" name="split-info"><DataField dataType="string" name="Season" optype="categorical"><Value value="fall"/></DataField></Extension></MiningBuildTask><DataDictionary
  numberOfFields="8">
  <DataField dataType="double" displayName="AODc" name="AODc"
optype="continuous"/>
  <DataField dataType="double" displayName="HPBL" name="HPBL"
optype="continuous"/>
  <DataField dataType="double" displayName="PM25" name="PM25"
optype="continuous"/>
```
<DataField dataType="double" displayName="PRCP" name="PRCP" optype="continuous"/>
<DataField dataType="double" displayName="VIS" name="VIS" optype="continuous"/>
<DataField dataType="double" displayName="uwnd" name="uwnd" optype="continuous"/>
<DataField dataType="double" displayName="vwnd" name="vwnd" optype="continuous"/>
<DataField dataType="double" displayName="ws" name="ws" optype="continuous"/>
</DataDictionary><TransformationDictionary>
<DerivedField dataType="double" name="AODcNorm" optype="continuous"/>
    <NormContinuous field="AODc">
        <LinearNorm norm="-1.30314251767485" orig="-0.0555552466452275"/>
        <LinearNorm norm="3.07218526375298" orig="0.534076559413047"/>
    </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="HPBLNorm" optype="continuous"/>
    <NormContinuous field="HPBL">
        <LinearNorm norm="-1.88562195809867" orig="264.5"/>
        <LinearNorm norm="3.14811881012861" orig="1580.564331"/>
    </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="PM25Norm" optype="continuous"/>
    <NormContinuous field="PM25">
        <LinearNorm norm="-1.40861165269159" orig="5.020833"/>
        <LinearNorm norm="3.58667346855772" orig="44.175"/>
    </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="PRCPNorm" optype="continuous"/>
    <NormContinuous field="PRCP">
        <LinearNorm norm="-0.297894048762357" orig="0"/>
        <LinearNorm norm="6.55185633008161" orig="1.09"/>
    </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="VISNorm" optype="continuous">
  <NormContinuous field="VIS">
    <LinearNorm norm="-2.6124648229923" orig="0.7889"/>
    <LinearNorm norm="0.857736384408414" orig="1.61"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="uwndNorm" optype="continuous">
  <NormContinuous field="uwnd">
    <LinearNorm norm="-1.87553795410954" orig="-4.413"/>
    <LinearNorm norm="2.40179960011494" orig="6.402"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="vwndNorm" optype="continuous">
  <NormContinuous field="vwnd">
    <LinearNorm norm="-2.73144872741407" orig="-6.413"/>
    <LinearNorm norm="2.72965771445759" orig="7.399"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="wsNorm" optype="continuous">
  <NormContinuous field="ws">
    <LinearNorm norm="-1.75454533697452" orig="0.1491876"/>
    <LinearNorm norm="4.09334172589358" orig="1.440432"/>
  </NormContinuous>
</DerivedField>

</TransformationDictionary>
<NeuralNetwork activationFunction="tanh" algorithmName="MLP" functionName="regression"><Extension extender="spss.com" name="modelID" value="7"/></NeuralNetwork>

<Extension extender="spss.com" name="modelID" value="7"/>
<MiningField name="PM25" usageType="predicted"/>
</MiningSchema><ModelExplanation><PredictiveModelQuality dataUsage="training" meanSquaredError="27.4775632488956" r-squared="0.495378467831876"
<DerivedField dataType="double" optype="continuous">
  <FieldRef field="VISNorm"/>
</DerivedField>
</NeuralInput>
<NeuralInput id="4">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="uwndNorm"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="5">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="vwndNorm"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="6">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="wsNorm"/>
  </DerivedField>
</NeuralInput>
</NeuralInputs>
<NeuralLayer numberOfNeurons="2">
  <Neuron bias="0.2234220817867" id="7">
    <Con from="0" weight="1.16158742355143"/>
    <Con from="1" weight="0.18519921174986"/>
    <Con from="2" weight="0.0640542742228049"/>
    <Con from="3" weight="-0.079656384233998"/>
    <Con from="4" weight="-0.177320570124772"/>
    <Con from="5" weight="0.125926586801668"/>
    <Con from="6" weight="-0.584952212602265"/>
  </Neuron>
  <Neuron bias="-0.761157065148482" id="8">
    <Con from="0" weight="0.225661593314747"/>
    <Con from="1" weight="-0.21898688814196"/>
    <Con from="2" weight="0.36732298151931"/>
    <Con from="3" weight="0.492389977802841"/>
    <Con from="4" weight="-0.0399479526379167"/>
    <Con from="5" weight="-0.15607732610719"/>
    <Con from="6" weight="-0.268390376392184"/>
  </Neuron>
</NeuralLayer>
<NeuralLayer activationFunction="identity" numberOfNeurons="1">
  <Neuron bias="-0.383906769312137" id="9">
    <Con from="7" weight="0.852144468828465"/>
    <Con from="8" weight="-0.670124913483496"/>
  </Neuron>
</NeuralLayer>
Cleveland

Figure C5. Neural network used for Cleveland fall season

Neural network code – fall

<?xml version="1.0" encoding="UTF-8"?>
<PMML version="4.1"
xmlns="http://www.dmg.org/PMML-4_1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema">
<Header copyright="(C) Copyright IBM Corp. 1994, 2013">
  <Application name="IBM SPSS Modeler Common" version="16.0.0.0"/>
</Header>

</PMML>
<Timestamp>Sat Jul 05 12:53:07 2014</Timestamp>
</Header><MiningBuildTask><Extension extender="spss.com" name="split-info"><DataField dataType="string" name="Season" optype="categorical"><Value value="fall"/></DataField></Extension></MiningBuildTask><DataDictionary numberOfFields="8">
<DataField dataType="double" displayName="AOD_total" name="AOD_total" optype="continuous"/>
<DataField dataType="string" displayName="AQI_Class" name="AQI_Class" optype="categorical">
  <Value value="Good"/>
  <Value value="Moderate"/>
</DataField>
<DataField dataType="double" displayName="PM25" name="PM25" optype="continuous"/>
<DataField dataType="double" displayName="airTemp" name="airTemp" optype="continuous"/>
<DataField dataType="double" displayName="hpbl_m" name="hpbl_m" optype="continuous"/>
<DataField dataType="double" displayName="rhum" name="rhum" optype="continuous"/>
<DataField dataType="double" displayName="vwnd" name="vwnd" optype="continuous"/>
<DataField dataType="double" displayName="ws" name="ws" optype="continuous"/>
</DataDictionary><TransformationDictionary>
<DerivedField dataType="double" name="AOD_totalNorm" optype="continuous">
  <NormContinuous field="AOD_total">
    <LinearNorm norm="-1.29729578054712" orig="-0.022000002"/>
    <LinearNorm norm="3.15963665046224" orig="0.505"/>
  </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="AQI_ClassValue0" optype="categorical">
  <NormDiscrete field="AQI_Class" value="Good"/>
</DerivedField>
<DerivedField dataType="double" name="AQI_ClassValue1" optype="categorical">
  <NormDiscrete field="AQI_Class" value="Moderate"/>
</DerivedField>
<DerivedField dataType="double" name="PM25Norm" optype="continuous">
  <NormContinuous field="PM25">
  </NormContinuous>
</DerivedField>
</TransformationDictionary>
<DerivedField dataType="double" name="airTempNorm" optype="continuous">
  <NormContinuous field="airTemp">
    <LinearNorm norm="-1.85374966492692" orig="273"/>
    <LinearNorm norm="1.86910509005434" orig="298.9668274"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="hpbl_mNorm" optype="continuous">
  <NormContinuous field="hpbl_m">
    <LinearNorm norm="-1.48655686275536" orig="299.2"/>
    <LinearNorm norm="2.99702539404228" orig="1718.2"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="rhumNorm" optype="continuous">
  <NormContinuous field="rhum">
    <LinearNorm norm="-2.37050403670295" orig="53"/>
    <LinearNorm norm="2.06284437690099" orig="85"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="vwndNorm" optype="continuous">
  <NormContinuous field="vwnd">
    <LinearNorm norm="-1.72651716005916" orig="-2.87"/>
    <LinearNorm norm="2.44792371484961" orig="7.79"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="wsNorm" optype="continuous">
  <NormContinuous field="ws">
    <LinearNorm norm="-1.15518480433227" orig="0.30864"/>
    <LinearNorm norm="3.45270117456198" orig="1.538056"/>
  </NormContinuous>
</DerivedField>
<TransformationDictionary><NeuralNetwork activationFunction="tanh"
algorithmName="MLP" functionName="regression"><Extension extender="spss.com"
name="modelID" value="1"/></NeuralNetwork></TransformationDictionary>

<MiningSchema>

<MiningField importance="0.224854395331076" name="AOD_total"/>

<MiningField importance="0.269804167338796" name="AQI_Class"/>

<MiningField importance="0.0239536338568436" name="airTemp"/>

<MiningField importance="0.251683106027131" name="hpbl_m"/>

<MiningField importance="0.0273966699293797" name="rhum"/>

<MiningField importance="0.044938550416736" name="vwnd"/>

<MiningField importance="0.157369477100038" name="ws"/>

<MiningField name="PM25" usageType="predicted"/></MiningSchema>

<ModelExplanation><PredictiveModelQuality dataUsage="training" meanSquaredError="46.383399466347" r-
squared="0.593189927680208" targetField="PM25"><LiftData><ModelLiftGraph><LiftGraph><XCoordinates><Array
n="21" type="int">1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58
69</Array></XCoordinates><YCoordinates><Array n="21" type="real">22.2625
127.6473333 67.3207 101.2238 85.2652 69.3625 68.3796 78.091 89.62086667
55.67076667 64.1999 83.978 58.94831765 53.908 39.06246667 29.8668 28.5835
31.1443 32.6653 28.0194667</Array></YCoordinates><BoundaryValues><Array n="21"
type="real">28.5160964187334 27.7368990046545 27.1868379031841 26.510627749614
25.8327061862551 24.423389327924 23.3676128123927 21.876323059366
20.5233600172514 17.8712011273896 14.0450645399302 12.4889782265677
11.6245075566088 10.9292877024035 10.41703282695 9.85165818519528
9.09211953952318 6.7649045091348</Array></BoundaryValues><BoundaryValueMeans><Array n="21"
type="real">29.7745395225606 28.6841553249082 27.9562419807011 27.2773912401752
26.828351002993 25.9370177664995 24.8526359635519 23.8016143684664
22.1904180153356 20.7056238092601 18.9868293416658 15.3008564701759
12.6928952724378 11.8307143538345 11.2797412505047 10.5440710159743
9.95095925396936 9.45660202195175 8.08275981374838</Array></BoundaryValueMeans></LiftGraph></ModelLiftGraph><LiftGraph><XCoordinates><Array
n="21" type="int">1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
</Array></XCoordinates><YCoordinates><Array n="21" type="real">30.4472897042832 29.928048183636 29.719243546718
28.5160964187334 27.7368990046545 27.1868379031841 26.510627749614
25.8327061862551 24.423389327924 23.3676128123927 21.876323059366
20.5233600172514 17.8712011273896 14.0450645399302 12.4889782265677
11.6245075566088 10.9292877024035 10.41703282695 9.85165818519528
9.09211953952318 6.7649045091348</Array></YCoordinates><BoundaryValues><Array n="21"
type="real">29.928048183636 29.719243546718 28.5160964187334 27.7368990046545
27.1868379031841 26.510627749614 25.8327061862551 24.423389327924
23.3676128123927 21.876323059366 20.5233600172514 17.8712011273896
14.0450645399302 12.4889782265677 11.6245075566088 10.9292877024035
10.41703282695 9.85165818519528 9.09211953952318 6.7649045091348</Array></BoundaryValues><BoundaryValueMeans><Array n="21"
type="real">29.7745395225606 28.6841553249082 27.9562419807011 27.2773912401752
26.828351002993 25.9370177664995 24.8526359635519 23.8016143684664
22.1904180153356 20.7056238092601 18.9868293416658 15.3008564701759
12.6928952724378 11.8307143538345 11.2797412505047 10.5440710159743
9.95095925396936 9.45660202195175 8.08275981374838</Array></BoundaryValueMeans></LiftGraph></ModelLiftGraph><PredictiveModelQuality dataUsage="test" meanSquaredError="31.6407062518601" r-
squared="0.613780868439043" targetField="PM25"></PredictiveModelQuality></ModelExplanation>
<FieldRef field="AOD_totalNorm"/>
</DerivedField>
</NeuralInput>
<NeuralInput id="1">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="AQI_ClassValue0"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="2">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="AQI_ClassValue1"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="3">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="airTempNorm"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="4">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="hpbl_mNorm"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="5">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="rhumNorm"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="6">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="vwndNorm"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="7">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="wsNorm"/>
  </DerivedField>
</NeuralInput>
</NeuralInputs>
<NeuralLayer numberOfNeurons="4">
  <Neuron bias="0.232254531457255" id="8">
    <Con from="0" weight="0.0754114132216524"/>
    <Con from="1" weight="-0.659993650095767"/>
    <Con from="2" weight="0.31288303101568"/>
    <Con from="3" weight="-0.105513063679318"/>
  </Neuron>
</NeuralLayer>
</NeuralLayer>
Figure C6. Neural network used for Cleveland winter season

Neural network code – winter

```xml
<?xml version="1.0" encoding="UTF-8"?><PMML version="4.1"
xmns="http://www.dmg.org/PMML-4_1"
xmns:xsi="http://www.w3.org/2001/XMLSchema"><Header copyright="(C) Copyright IBM Corp. 1994, 2013">
  <Application name="IBM SPSS Modeler Common" version="16.0.0.0"/>
  <Timestamp>Sat Jul 05 12:53:07 2014</Timestamp>
</Header><MiningBuildTask><Extension extender="spss.com" name="split-info"><DataField dataType="string" name="Season" optype="categorical"><Value value="Winter"/></DataField></Extension></MiningBuildTask><DataDictionary numberOfFields="8">
```
<DataField dataType="double" displayName="AOD_total" name="AOD_total" optype="continuous"/>
<DataField dataType="string" displayName="AQI_Class" name="AQI_Class" optype="categorical">
    <Value value="Good"/>
    <Value value="Moderate"/>
</DataField>
<DataField dataType="double" displayName="PM25" name="PM25" optype="continuous"/>
<DataField dataType="double" displayName="airTemp" name="airTemp" optype="continuous"/>
<DataField dataType="double" displayName="hpbl_m" name="hpbl_m" optype="continuous"/>
<DataField dataType="double" displayName="rhum" name="rhum" optype="continuous"/>
<DataField dataType="double" displayName="vwnd" name="vwnd" optype="continuous"/>
<DataField dataType="double" displayName="ws" name="ws" optype="continuous"/>
</DataDictionary><TransformationDictionary>
    <DerivedField dataType="double" name="AOD_totalNorm" optype="continuous">
        <NormContinuous field="AOD_total">
            <LinearNorm norm="-0.980121427490505" orig="0.001"/>
            <LinearNorm norm="3.55709116718582" orig="0.798"/>
        </NormContinuous>
    </DerivedField>
    <DerivedField dataType="double" name="AQI_ClassValue0" optype="categorical">
        <NormDiscrete field="AQI_Class" value="Good"/>
    </DerivedField>
    <DerivedField dataType="double" name="AQI_ClassValue1" optype="categorical">
        <NormDiscrete field="AQI_Class" value="Moderate"/>
    </DerivedField>
    <DerivedField dataType="double" name="PM25Norm" optype="continuous">
        <NormContinuous field="PM25">
            <LinearNorm norm="-1.52320184374666" orig="4.0125"/>
            <LinearNorm norm="1.76713654210638" orig="25.9083"/>
        </NormContinuous>
    </DerivedField>
</TransformationDictionary>
<DerivedField data-type="double" name="airTempNorm" optype="continuous">
    <NormContinuous field="airTemp">
        <LinearNorm norm="-2.46949258200547" orig="264.3"/>
        <LinearNorm norm="2.5514776754743" orig="286"/>
    </NormContinuous>
</DerivedField>

<DerivedField data-type="double" name="hpbl_mNorm" optype="continuous">
    <NormContinuous field="hpbl_m">
        <LinearNorm norm="-1.36775135929833" orig="363.6902"/>
        <LinearNorm norm="1.74927792735219" orig="1538.2"/>
    </NormContinuous>
</DerivedField>

<DerivedField data-type="double" name="rhumNorm" optype="continuous">
    <NormContinuous field="rhum">
        <LinearNorm norm="-1.92821171534781" orig="65"/>
        <LinearNorm norm="1.2703315161721" orig="91"/>
    </NormContinuous>
</DerivedField>

<DerivedField data-type="double" name="vwndNorm" optype="continuous">
    <NormContinuous field="vwnd">
        <LinearNorm norm="-1.94187486239973" orig="-2.23"/>
        <LinearNorm norm="1.8649927839815" orig="9.39"/>
    </NormContinuous>
</DerivedField>

<DerivedField data-type="double" name="wsNorm" optype="continuous">
    <NormContinuous field="ws">
        <LinearNorm norm="-1.11315396950661" orig="0.509256"/>
        <LinearNorm norm="2.51515996470348" orig="2.00616"/>
    </NormContinuous>
</DerivedField>

</TransformationDictionary>
<NeuralNetwork activationFunction="tanh" algorithmName="MLP" functionName="regression"><Extension extender="spss.com" name="modelID" value="0"></MiningSchema>

<MiningField importance="0.0458950771906677" name="AOD_total"/>
<MiningField name="AQI_Class"/>
<MiningField name="airTemp"/>
<MiningField name="hpbl_m"/>
<MiningField name="rhum"/>
<MiningField name="vwnd"/>
<MiningField name="ws"/>
<MiningField name="PM25" usageType="predicted"/>

dataUsage="training" meanSquaredError="3.14823502523871" r-squared="0.918911097970901"
targetField="PM25">

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Boundary Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.47083333</td>
<td>22.2469872854479</td>
</tr>
<tr>
<td>2</td>
<td>33.5783</td>
<td>22.2263713706615</td>
</tr>
<tr>
<td>3</td>
<td>16.316</td>
<td>22.0061197839418</td>
</tr>
<tr>
<td>4</td>
<td>28.088</td>
<td>21.8579734466412</td>
</tr>
<tr>
<td>5</td>
<td>18.228</td>
<td>20.896740735632</td>
</tr>
<tr>
<td>6</td>
<td>22.75</td>
<td>20.29225355585472</td>
</tr>
<tr>
<td>7</td>
<td>17.864</td>
<td>19.6678198020342</td>
</tr>
<tr>
<td>8</td>
<td>24.766666667</td>
<td>14.9575765628959</td>
</tr>
<tr>
<td>9</td>
<td>17.264</td>
<td>13.5663452899922</td>
</tr>
<tr>
<td>10</td>
<td>14.475</td>
<td>13.599836317299</td>
</tr>
<tr>
<td>11</td>
<td>14.5375</td>
<td>11.0549443277648</td>
</tr>
<tr>
<td>12</td>
<td>17.0167</td>
<td>9.57033713643319</td>
</tr>
<tr>
<td>13</td>
<td>9.2125</td>
<td>9.20770492655235</td>
</tr>
<tr>
<td>14</td>
<td>8.0792</td>
<td>8.91230737855854</td>
</tr>
<tr>
<td>15</td>
<td>4.1458</td>
<td>8.8505754823881</td>
</tr>
<tr>
<td>16</td>
<td>6.25</td>
<td>8.63336345173837</td>
</tr>
<tr>
<td>17</td>
<td>6.2791667</td>
<td>8.526279949808</td>
</tr>
<tr>
<td>18</td>
<td>9.99374215207709</td>
<td>8.1138794757209</td>
</tr>
<tr>
<td>19</td>
<td>14.9575765628959</td>
<td>5.71461653433884</td>
</tr>
<tr>
<td>20</td>
<td>13.5663452899922</td>
<td>2.2469872854479</td>
</tr>
</tbody>
</table>

**Mean Squared Error:** 34.427325255276

**R-squared:** 0.266672208030013
<Con from="0" weight="-0.36287253814356"/>
<Con from="1" weight="0.223370847185214"/>
<Con from="2" weight="-0.0633565560643811"/>
<Con from="3" weight="-0.348263492023647"/>
<Con from="4" weight="0.459056789669606"/>
<Con from="5" weight="-0.144597198838542"/>
<Con from="6" weight="0.442619341182429"/>
<Con from="7" weight="0.091204106203013"/>
</Neuron>
</NeuralLayer><NeuralLayer activationFunction="identity" numberOfNeurons="1">
  <Neuron bias="0.215912727396739" id="11">
    <Con from="8" weight="1.33044310015913"/>
    <Con from="9" weight="-0.334317410692025"/>
    <Con from="10" weight="0.04710679679262"/>
  </Neuron>
</NeuralLayer><NeuralOutputs>
  <NeuralOutput outputNeuron="11">
    <DerivedField dataType="double" optype="continuous">
      <FieldRef field="PM25Norm"/>
    </DerivedField>
  </NeuralOutput>
</NeuralOutputs></NeuralNetwork></PMML>
Figure C7. Neural network used for Cleveland summer season

**Neural network code – summer**

```xml
<?xml version="1.0" encoding="UTF-8"?>
<PMML version="4.1"
xmlns="http://www.dmg.org/PMML-4_1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema">
    <Header copyright="(C) Copyright IBM Corp. 1994, 2013">
        <Application name="IBM SPSS Modeler Common" version="16.0.0.0"/>
        <Timestamp>Sat Jul 05 12:53:07 2014</Timestamp>
    </Header>
    <MiningBuildTask>
        <Extension extender="spss.com" name="split-info">
            <DataField dataType="string" name="Season" optype="categorical">
                <Value value="Summer"/>
            </DataField>
        </Extension>
        <DataDictionary numberOfFields="8">
            <DataField dataType="double" displayName="AOD_total" name="AOD_total" optype="continuous"/>
            <DataField dataType="string" displayName="AQI_Class" name="AQI_Class" optype="categorical">
                <Value value="Good"/>
                <Value value="Moderate"/>
            </DataField>
        </DataDictionary>
    </MiningBuildTask>
</PMML>
```
<DerivedField dataType="double" name="hpbl_mNorm" optype="continuous">
    <NormContinuous field="hpbl_m">
        <LinearNorm norm="-1.4860826986479" orig="453.6"/>
        <LinearNorm norm="3.64017029493377" orig="1586.4167"/>
    </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="rhumNorm" optype="continuous">
    <NormContinuous field="rhum">
        <LinearNorm norm="-2.95741322464574" orig="49.21246"/>
        <LinearNorm norm="1.86234539693322" orig="87"/>
    </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="vwndNorm" optype="continuous">
    <NormContinuous field="vwnd">
        <LinearNorm norm="-2.01583934485319" orig="-3.88"/>
        <LinearNorm norm="2.62811805686653" orig="6.99"/>
    </NormContinuous>
</DerivedField>

<DerivedField dataType="double" name="wsNorm" optype="continuous">
    <NormContinuous field="ws">
        <LinearNorm norm="-1.37489084727064" orig="0.36008"/>
        <LinearNorm norm="3.76361070209842" orig="1.538056"/>
    </NormContinuous>
</DerivedField>

</TransformationDictionary>
</NeuralNetwork activationFunction="tanh" algorithmName="MLP" functionName="regression">
    <Extension extender="spss.com" name="modelID" value="3"/>
    <MiningSchema>
        <MiningField importance="0.0694640575450446" name="AOD_total"/>
        <MiningField importance="0.173154378690406" name="AQI_Class"/>
        <MiningField importance="0.13576656960779" name="airTemp"/>
        <MiningField importance="0.162751111782125" name="hpbl_m"/>
    </MiningSchema>
</NeuralNetwork>
<table>
<thead>
<tr>
<th>X Coordinates</th>
<th>Y Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 8 14 20 26 32 38 44 50 56 62 68 74 80 86 92 104 110 116 120</td>
<td>66.884 187.7692 172.544988 174.0903 144.02758985 150.824 150.98440884 145.70488333 121.58533333 139.69189523 110.7623333 81.34345 89.96319341 71.0339 64.8668</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Boundary Values</th>
<th>Boundary Value Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.255980944175</td>
<td>29.3532182537495</td>
</tr>
</tbody>
</table>

**Model Explanation**

<table>
<thead>
<tr>
<th>Neural Inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DerivedField dataType=&quot;double&quot; optype=&quot;continuous&quot;</td>
</tr>
<tr>
<td>FieldRef field=&quot;AOD_totalNorm&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neural Inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DerivedField dataType=&quot;double&quot; optype=&quot;continuous&quot;</td>
</tr>
<tr>
<td>FieldRef field=&quot;AQI_ClassValue0&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neural Inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DerivedField dataType=&quot;double&quot; optype=&quot;continuous&quot;</td>
</tr>
<tr>
<td>FieldRef field=&quot;AQI_ClassValue1&quot;</td>
</tr>
</tbody>
</table>
<DerivedField/>
</NeuralInput>
<NeuralInput id="3">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="airTempNorm"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="4">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="hpbl_mNorm"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="5">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="rhumNorm"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="6">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="vwndNorm"/>
  </DerivedField>
</NeuralInput>
<NeuralInput id="7">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="wsNorm"/>
  </DerivedField>
</NeuralInput>
</NeuralInputs>
<NeuralLayer numberOfNeurons="6">
  <Neuron bias="-0.0110578360176585" id="8">
    <Con from="0" weight="0.0530444944937063"/>
    <Con from="1" weight="-0.65499529094326"/>
    <Con from="2" weight="0.169846516003584"/>
    <Con from="3" weight="0.085460050114479"/>
    <Con from="4" weight="-0.128876586135986"/>
    <Con from="5" weight="-0.0785654522922671"/>
    <Con from="6" weight="-0.291415419045591"/>
    <Con from="7" weight="-0.300691735013248"/>
  </Neuron>
  <Neuron bias="-0.391513074122826" id="9">
    <Con from="0" weight="0.17036544494397063"/>
    <Con from="1" weight="0.474813799310407"/>
    <Con from="2" weight="-0.350841700341941"/>
    <Con from="3" weight="0.291376408685262"/>
    <Con from="4" weight="0.156870560721686"/>
</NeuralLayer><NeuralLayer activationFunction="identity" numberOfNeurons="1">
    <Neuron bias="-0.036334985345291" id="14">
        <Con from="8" weight="0.517230269780903"/>
        <Con from="9" weight="-0.248427294675024"/>
        <Con from="10" weight="-0.453322788954358"/>
        <Con from="11" weight="0.15557322374047"/>
        <Con from="12" weight="-0.412379548424647"/>
        <Con from="13" weight="-0.357163453256766"/>
    </Neuron>
</NeuralLayer><NeuralOutputs>
    <NeuralOutput outputNeuron="14">
        <DerivedField dataType="double" optype="continuous">
            <FieldRef field="PM25Norm"/>
        </DerivedField>
    </NeuralOutput>
</NeuralOutputs></NeuralNetwork></PMML>
Figure C8. Neural network used for Cleveland spring season

Neural network code – Spring

```xml
<?xml version="1.0" encoding="UTF-8"?><PMML version="4.1"
xmlns="http://www.dmg.org/PMML-4_1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema"/>

<Header copyright="(C) Copyright IBM Corp. 1994, 2013">
  <Application name="IBM SPSS Modeler Common" version="16.0.0.0"/>
  <Timestamp>Sat Jul 05 12:53:07 2014</Timestamp>
</Header>

<MiningBuildTask>
  <Extension extender="spss.com" name="split-info">
    <DataField dataType="string" name="Season" optype="categorical">
      <Value value="Spring"/>
    </DataField>
  </Extension>
</MiningBuildTask>

<DataDictionary numberOfFields="8">
  <DataField dataType="double" displayName="AOD_total" name="AOD_total" optype="continuous"/>
  <DataField dataType="string" displayName="AQI_Class" name="AQI_Class" optype="categorical">
    <Value value="Good"/>
    <Value value="Moderate"/>
  </DataField>
</DataDictionary>
```
<DerivedField>
<DerivedField dataType="double" name="hpbl_mNorm"
optype="continuous">
  <NormContinuous field="hpbl_m">
    <LinearNorm norm="-1.5125473175656" orig="310.3"/>
    <LinearNorm norm="3.39153912485866" orig="1753.1"/>
  </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="rhumNorm"
optype="continuous">
  <NormContinuous field="rhum">
    <LinearNorm norm="-2.28676770068135" orig="48"/>
    <LinearNorm norm="2.04767398660817" orig="90"/>
  </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="vwndNorm"
optype="continuous">
  <NormContinuous field="vwnd">
    <LinearNorm norm="-2.8422577488717" orig="-9.54"/>
    <LinearNorm norm="2.1738709197737" orig="8.66"/>
  </NormContinuous>
</DerivedField>
<DerivedField dataType="double" name="wsNorm"
optype="continuous">
  <NormContinuous field="ws">
    <LinearNorm norm="-1.47322277293251" orig="0.41152"/>
    <LinearNorm norm="2.92818968736494" orig="1.44032"/>
  </NormContinuous>
</DerivedField>
</TransformationDictionary>
</NeuralNetwork>

<Extension extender="spss.com" name="modelID" value="2"/>

<MiningSchema>
  <MiningField importance="0.109096539631396" name="AOD_total"/>
  <MiningField importance="0.201203668412377" name="AQL_Class"/>
  <MiningField importance="0.292863059970078" name="airTemp"/>
  <MiningField importance="0.0981913552868438" name="hpbl_m"/>
  <MiningField importance="0.118963905563436" name="rhum"/>
19.7868039211961 19.2256613702292 18.9671516371893 17.135147199498
15.3236982456364 13.3575973915877 11.1618936059801 10.6164781640482
9.26442972346409

ModelExplanation

NeuralInputs

<NeuralInput id="0">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="AOD_totalNorm"/>
    </DerivedField>
  </NeuralInput>

<NeuralInput id="1">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="AQI_ClassValue0"/>
    </DerivedField>
  </NeuralInput>

<NeuralInput id="2">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="AQI_ClassValue1"/>
    </DerivedField>
  </NeuralInput>

<NeuralInput id="3">
  <DerivedField dataType="double" optype="continuous">
    <FieldRef field="AQI_ClassValue2"/>
    </DerivedField>
  </NeuralInput>
<DerivedField dataType="double" optype="continuous">
  <FieldRef field="airTempNorm"/>
</DerivedField>
</NeuralInput>
</NeuralInputs>
<NeuralLayer numberOfNeurons="5">
  <Neuron bias="0.39839345938855" id="8">
    <Con from="0" weight="0.150843638934153"/>
    <Con from="1" weight="0.0803715148090612"/>
    <Con from="2" weight="-0.0892505679793591"/>
    <Con from="3" weight="0.323247905547716"/>
    <Con from="4" weight="0.335528022773959"/>
    <Con from="5" weight="0.307974639746435"/>
    <Con from="6" weight="-0.221850117228355"/>
    <Con from="7" weight="-0.434776672701836"/>
  </Neuron>
  <Neuron bias="0.451781900776299" id="9">
    <Con from="0" weight="-0.10542312928365"/>
    <Con from="1" weight="-0.411892894915284"/>
    <Con from="2" weight="-0.204681331786599"/>
    <Con from="3" weight="-0.153747036505367"/>
    <Con from="4" weight="0.366973950341058"/>
    <Con from="5" weight="-0.445141656735156"/>
    <Con from="6" weight="0.0251828942141805"/>
    <Con from="7" weight="-0.0953405546469603"/>
  </Neuron>
</NeuralLayer>
<Neuron.bias="0.0575971536991476" id="10">
  <Con.from="0" weight="0.0278406818815677"/>
  <Con.from="1" weight="-0.357088452520013"/>
  <Con.from="2" weight="-0.26218589562149"/>
  <Con.from="3" weight="0.343611971544421"/>
  <Con.from="4" weight="-0.0988542201055472"/>
  <Con.from="5" weight="-0.214232489352707"/>
  <Con.from="6" weight="-0.170097841541144"/>
  <Con.from="7" weight="0.303041933600693"/>
</Neuron>

<Neuron.bias="-0.756001777808766" id="11">
  <Con.from="0" weight="0.00136759378995834"/>
  <Con.from="1" weight="-1.075209686767973"/>
  <Con.from="2" weight="-0.331798968433764"/>
  <Con.from="3" weight="0.748550090211983"/>
  <Con.from="4" weight="-0.274089499349548"/>
  <Con.from="5" weight="-0.253367994362143"/>
  <Con.from="6" weight="0.251519370555115"/>
  <Con.from="7" weight="0.099163722194529"/>
</Neuron>

<Neuron.bias="-0.29702783264711" id="12">
  <Con.from="0" weight="-0.404680945688334"/>
  <Con.from="1" weight="0.392673780570402"/>
  <Con.from="2" weight="-1.01976227001758"/>
  <Con.from="3" weight="0.0363055805689161"/>
  <Con.from="4" weight="-0.028563670602223"/>
  <Con.from="5" weight="-0.133979256251845"/>
  <Con.from="6" weight="-0.0913165887154628"/>
  <Con.from="7" weight="0.411156182209554"/>
</Neuron>

</NeuralLayer><NeuralLayer activationFunction="identity" numberOfNeurons="1">
  <Neuron.bias="0.371068793600078" id="13">
    <Con.from="8" weight="-0.0295741315805067"/>
    <Con.from="9" weight="0.0495884889075019"/>
    <Con.from="10" weight="0.260639363614453"/>
    <Con.from="11" weight="0.995782089229084"/>
    <Con.from="12" weight="-0.749429138228611"/>
  </Neuron>
</NeuralLayer><NeuralOutputs>
  <NeuralOutput outputNeuron="13">
    <DerivedField dataType="double" otype="continuous">
      <FieldRef field="PM25Norm"/>
    </DerivedField>
  </NeuralOutput>
</NeuralOutputs>
</DerivedField>
</NeuralOutput>
</NeuralOutputs></NeuralNetwork></PMML>