PROTECTING HUMAN HEALTH: FORECASTING PM FINE LEVELS FOR AIR QUALITY INDEX REPORTING

JASON P. BREWER, MATTHEW S. JOHNSON
UNDERGRADUATE METEOROLOGY MAJORS
AND
MICHAEL Y. THELEN
UNDERGRADUATE STATISTICS MAJOR

Abstract. Fine particulate matter is a significant pollutant that endangers human health. Small particulates, 2.5 micrometers in diameter or less, penetrate further into the lungs of humans than larger particulates leading to increased cases of respiratory diseases and eventual death. Both annual mean and 24 hour National Ambient Air Quality Standards have been set for fine particulate matter (PM2.5). PM2.5 is one of five pollutants reported in the USEPAs Air Quality Index. It is critically important that todays PM2.5 value can be accurately forecast so it can be reported to the public with an appropriate health advisory. Our objective is to develop reliable forecasting regression models to serve as tools for predicting PM2.5. The regression models will take into account various meteorological parameters such as temperature, wind speed, wind direction, and yesterdays PM2.5 measurements. Our client, the Maryland Department of Environment, provided all meteorological and particulate matter data. Analyses of selected particulate matter monitoring stations and meteorological sites in the state of Maryland have lead to discoveries of certain PM2.5 patterns. Trends show PM2.5 variations between winter and summer seasons as well as weekday and weekend periods. Various patterns, interaction terms, nonlinear curvature, and other possible confounders will be taken into account. Regression analysis and model building techniques will be implemented for prognostic purposes and also for interests in inferential procedures on linear combinations of regression variables. Development of more specific regression models and software packages for these different periods will improve future forecasts of PM2.5 in addition to making the information readily accessible to the public.

1. Introduction

Fine particulate matter (PM2.5) defined as ultra-fine masses of size $2.5(\mu/m^3)$ and less is one of six major criteria pollutants. The presence and prevalence of PM fine has effected the lives of many. Air quality index reporting has extended standards to maintain acceptable levels of PM fine based on studies in the chemical makeup and its health impacts.

1.1. **Chemistry.** PM2.5 is directly emitted from both anthropogenic and natural sources. Anthropogenic effects, processes, objects or materials are polluting derivatives from human

Date: 4.13.2007.

activities such as power plants, industry, transportation, mining, habitations and agriculture techniques and machinery. The release of gases and dust into the atmosphere due to waste disposal practices have been linked to increases in air and water pollution by a variety of studies. Additionally, slash-and-burn techniques, hydrological diversion, salinization and chemical pollutants from fertilizers and pesticides have plagued the air quality indexes linked to the agriculture industry. Air pollutants are commonly released as a result of mineral refining processes. Natural sources also exist as dirt, dust, sea salt and volcanic or fire ash. These sources are not derived from human activities and involvement but are still involved in various facets, such as the soiling effect, and contribute to the underlying issue of fine matter pollutants. Primary sources continue to be characterized from wind-blown dust and diesel exhaust, while secondary sources exist from the reaction of sulfates and nitrates released from power plants.

1.2. **Health Impacts.** Fine particles are inhalable and can penetrate into the lungs. Some are small enough to enter the bloodstream and then deposit themselves to be accumulated or absorbed by the body. Consistency across studies has shown tendencies of significant health risks such as respiratory and cardiovascular problems, coughing, painful breathing, asthma, bronchitis, emphysema, decreased lung function, weakening of the heart, heart attacks and premature death. Older women living in the most polluted cities have about a 150% increased risk of death from heart attacks related to particulate matter[1]. Recent studies have also underscored that the elderly with pre-existing cardiopulmonary disease are most at risk. In addition, other groups such as the very young, asthmatics, and diabetics may also be susceptible to the effects of PM. Increased mortality in infants and lung cancer patients from decreased lung function and inflammed airways has also exacerbated the prevalence of PM fine related diseases. Findings have suggested that extended exposure to PM can lead to chronic disease and/or a shortened life span[2].

1.3. **Purpose.** The purpose of our study is to develop models that will aid in the monitoring and prediction of fine particulate matter since PM fine is a major constituent of the Air Quality Index. However, due to air quality standards, a *Federal Reference Method*, or "PM_{FRM}," must be used, because it contains true information, by definition. Unfortunately, PM_{FRM} recorded data is not available in real-time, but instead either every three days or six days with a running midnight-to-midnight figure. Thus, the use of continuous PM fine measurements (PM_{cont}) will be implemented. These measurements are available electronically in real-time, on an hourly basis. The implication of a daily statistic has raised the issue of using the daily average PM_{cont} as a predictor for the federal reference method. We will show that the development of regression models to take daily average PM_{cont} readings and make it "PM_{FRM}-like" is effective. Further information, namely PM_{cont} measurements will be provided from air monitoring stations, such as National Air Monitoring Stations (NAMS),

[1]Study found in the New England Journal of Medicine

[2]According to EPA literature and publications

Airport Site	PM Site	Distance (miles)
KPHL (Philadelphia International Airport)	0003	≈ 35.24
KHGR (Hagerstown Washington Co. Airport)	0009	≈ 9.19
KDCA (Ronald Reagan Washington National Airport)	0030	≈ 16.45
KBWI (Baltimore Washington International Airport)	0040	≈ 10.39

TABLE 1. Meteorological conditions for airport and PM sites are assumed to be comparatively uniform, even across great distances.

Site	Data Percentage for 2001-06 Time Span
KBWI	≈ 98.22
KDCA	≈ 89.73
KHGR	≈ 99.22
KPHL	≈ 85.62

TABLE 2. Meteorological data completeness percentages.

State and Local Air Monitoring Stations (SLAMS), Special Purpose Monitors (SPMS), Photochemical Assessment Monitoring Stations (PAMS) and PM fine Chemical Speciation sites. Our final objective is to observe yearly trends and to develop functions in terms of yesterday's PM fine value and today's meteorological conditions to be used for forecasting purposes.

2. Preliminary Findings and Assumptions

PM_{FRM} data is collected by the Maryland Department of the Environment (MDE) on a filter over a 24-hour period and are measured every 3 or 6 days. Continuous hourly data from Tapered Elemental Oscillating Microbalances (TEOM) and Beta Attenuation Monitors (BAMM) will be the focus of this paper. Data from BAMM, FDMS, and TEOM sites are calculated on one-hour averages (60 1-minute averages, starting at minute 0 and ending at minute 59), without hour-to-hour overlap. Meteorological data are also available for these domains from local airports and measurement stations collected by the National Oceanic and Atmospheric Administration (NOAA) and were provided by MDE, all in an hourly form.

Five sites were found containing PM_{cont} readings; however, only four were co-located with PM_{FRM}. Co-location is necessary in developing a mapping model to transform PM_{cont} readings into PM_{FRM}-like measurements. Thus, site 0002 with only PM_{cont} was thrown out.

We created the air quality and meteorological dataset by merging the air monitoring site with the closest airport and the meteorological site. We assumed that meteorological conditions are comparatively uniform for our prospective regression analyses. We used the EPA quality assurance requirement of 75% completeness, or 18 hours, to define a valid day. Only valid days were included in this analysis.

The meteorological data collected at Baltimore-Washington International Airport was over 90% complete. In general, the meteorological dataset completeness was more than sufficient.

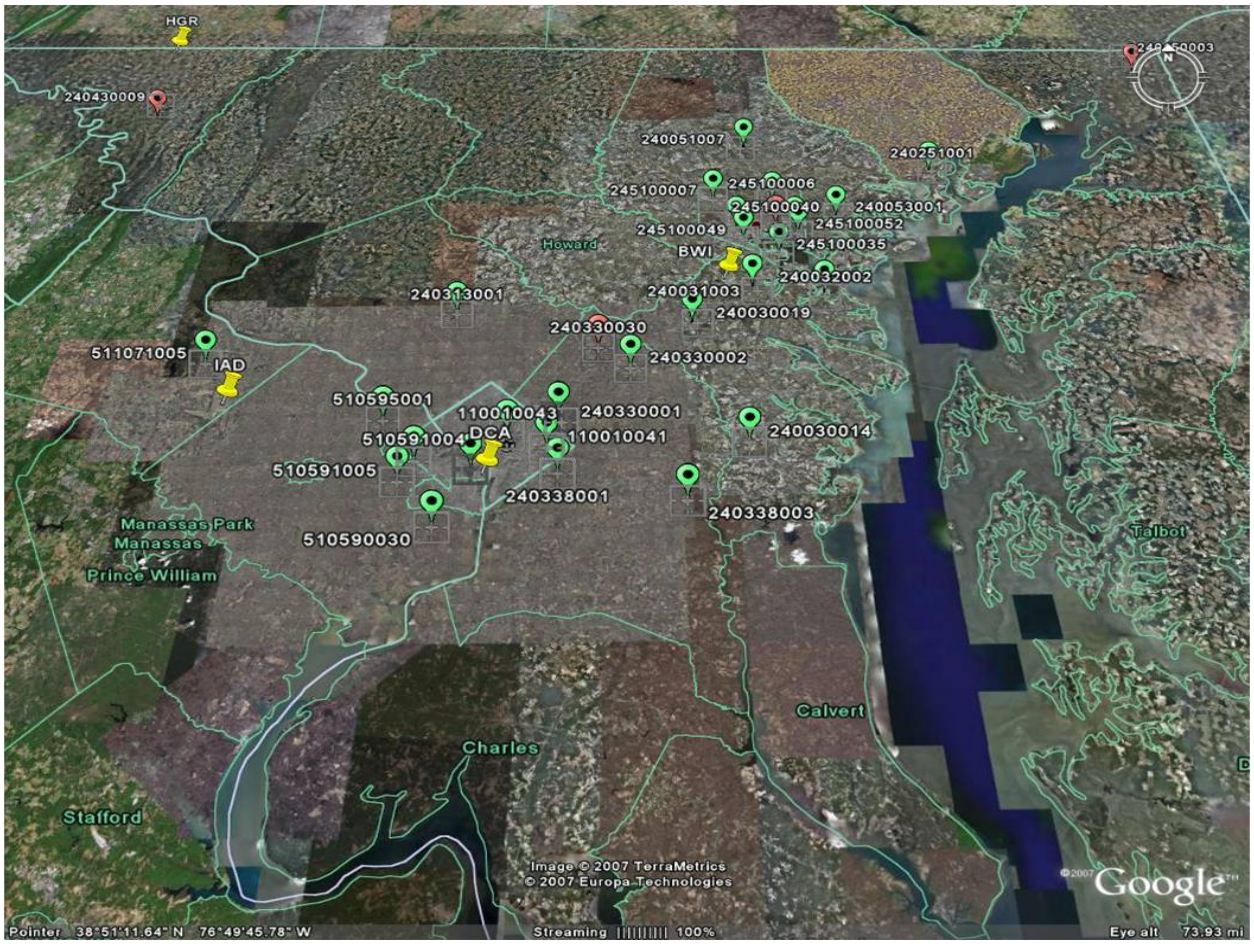

FIGURE 1. PM_{FRM}, PM_{cont} and airport meteorological site locations.

PM Site	County	Met Site	County
0003	Cecil	KPHL	Philadelphia[3]
0009	Washington	KHGR	Washington
0030	Prince George's	KDCA	Arlington[4]
0040	Baltimore (City)	KBWI	Anne Arundel

TABLE 3. County locations for airport and PM sites. Sites are located in the state of Maryland, with the exception of: [3]in Pennsylvania and [4]in Virginia.

Most air monitoring sites were located in different counties than the meterological data. In fact, only one pair out of four sites happens to be located in precisely the same county (Washington county, MD–see table 3).

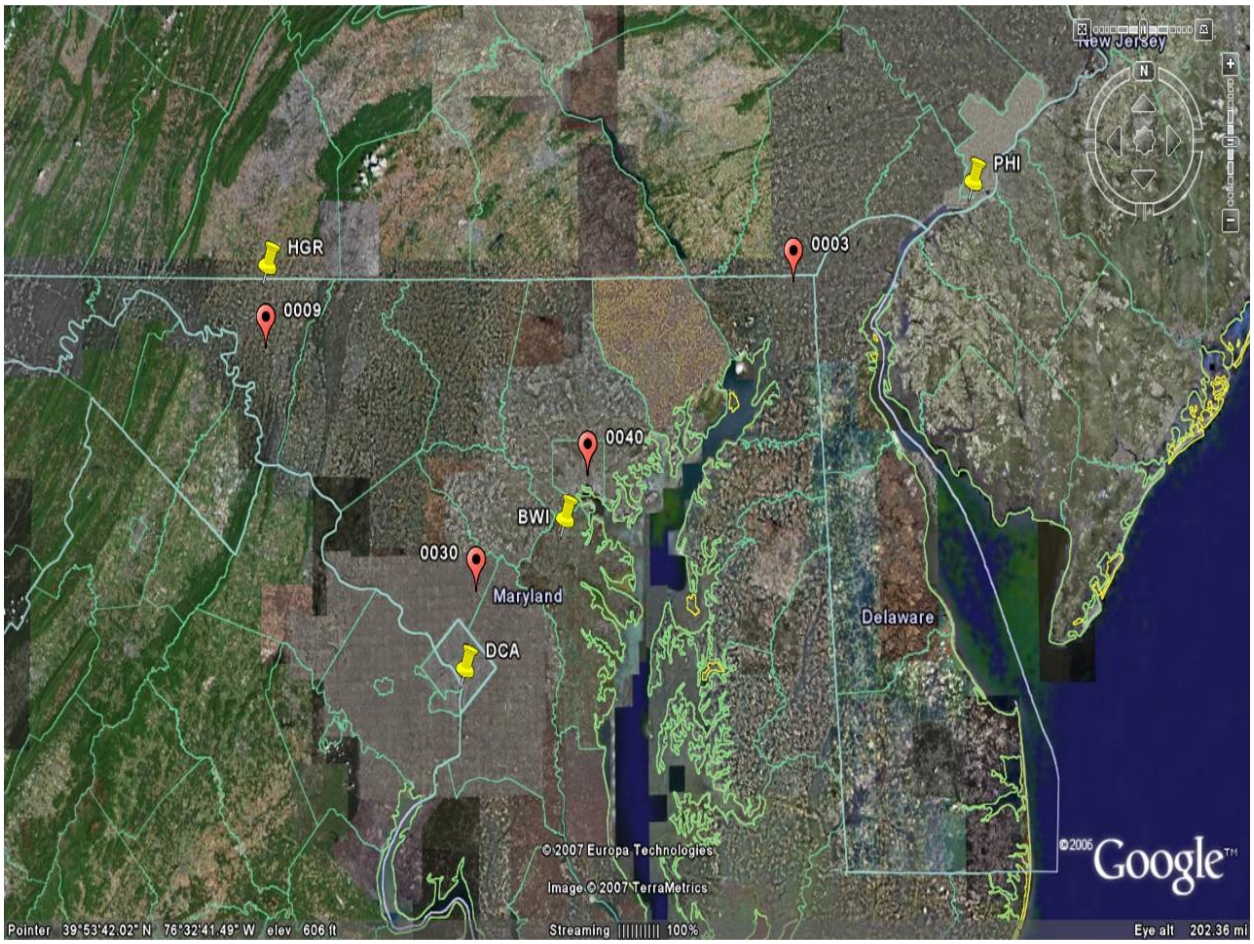

FIGURE 2. PM_{FRM}/PM_{cont} and airport meteorological sites. The development of a pairing scheme for nearly co-located sites relied heavily on latitude and longitudes, allow us to pair up sites in close proximity to each other.

Site	Approximate Data Percentage for 2001-06 Time Span	Date of First Observation
0003	≈ 4.34	6/28/06
0009	≈ 21.86	5/16/05
0030	≈ 13.65	9/15/05
0040	≈ 78.09	01/01/01

TABLE 4. PM_{cont} data completeness percentages. Notice only one site (0040) with data beginning from the start of the trend period.

Site	Approximate Dataset Percentages Prior to Merging	Approximate Final Dataset Percentages
PHL/0003	98.22/4.34	55.96/2.56
HGR/0009	89.73/21.86	61.11/14.88
DCA/0030	99.22/13.65	56.23/8.90
BWI/0040	85.62/78.09	57.46/44.50

TABLE 5. Approximate dataset percentage, prior to and following data merge. Final percentages reflect that of analyses directly preceding full-scale regression. Entries given in "Meteorological/PM" form.

In order to increase the amount of data available for analysis, data from four pairs of sites were used. Each air monitoring observation was matched with that of the nearest airport. Since we are focusing on a general model to forecast PM fine for the State of Maryland, we initially focused on all of the data up to the 95th percentile. It should be noted that a separate analysis with the full dataset was also performed, both sets yielding promising results.

It can be said that data completeness was affected by the start day for recorded PM_{cont} observations. In fact, most of the data lies within the years of 2005-06, with little to drive a regression analyses and exploratory study for the years 2001-04. In addition, assuming uniform meteorology across large distances may not accurately reflect what is happening meteorologically as we would like it to be at the air monitoring sites.

3. Regression Analyses

(Please see Appendix A for Regression Theory.) Our notion is to build a general piecewise model with stratifications, based on central PM_{cont} values. Implementing the Central Limit Theorem, we can note that the daily average is a well-behaved statistic with low variance. These averages can be easily regressed to a PM_{FRM} value after the predicted average has been computed. We conducted our analyses with two datasets: the full set, and the abbreviate set, which focuses on data up to the 95th percentile, to eliminate peak values, since averages are unstable with outliers.

Meteorological and PM_{cont} data are collected in an hourly fashion. Therefore, the two data sets could be merged and each hourly PM fine value was matched with its corresponding meteorological data. However, because the Air Quality Index uses a midnight-to-midnight measurement, a 24-hour average PM fine value was calculated for midnight-to-midnight. The daily average PM fine data were then compared with the daily meteorological variables. We were able to develop "level one" and "level two" explanatory variables to be implemented for our work. The following level one variables (and square terms for each, as indicated within the parentheses) were considered:

- Peak Temperature (X_6)
- Accumulated Precipitation (X_7)

- Maximum Wind Speed (X_8)
- Average Wind Speed (X_9)
- Average Pressure (X_{10})
- Average Humidity (X_{11})
- u vector (X_{12})
- v vector (X_{13})
- Yesterday's PM_{cont} value, "PM_{yes}" (X_{14})
- Probability of Precipitation (POP) (X_{15})

Arguable amounts of curvature were observed for each level one variable. Thus, it was natural to include square terms for each of these variables. In addition, interaction terms (falling under the category of level two explanatory variables) were also considered:

- $X1$ = (Maximum Wind Speed) $\times$ (Avg WS)
- $X2$ = (POP) $\times$ (Accumulated Precipitation)
- $X3$ = (POP) $\times$ (Average Humidity)
- $X4$ = (Accumulated Precipitation) $\times$ (Average Humidity)

One can consider our situation as potentially benefiting from a stratified piecewise modeling system. Such a system is resemblant of a two-way layout with two sub-factors: a seasonal effect (summer vs. winter)[5] and a weekly effect (weekend vs. weekday)[6]. This yields to one full model and eight additional models with some sort of stratification, for a total of nine models. The applicable output and SAS implementation is available (see last section for follow up information). Dummy variables were also considered on an exploratory basis; however, these variables were highly insignificant in any model and not considered for our final reports. Instead, a piecewise scenario is encouraged allowing square terms to account for additional curvature. Also, varying selections of meteorological variables were considered for each stratification. This approach will present increased R^2 values, thus increasing the ability to forecast.

4. Results

It can be seen that our R^2 values increase as stratifications become more specific with both the full and abbreviate dataset.

Using the abbreviated dataset, the following significant level one variables were found to be useful in prediction with:

Peak Temperature: Full, Summer, Weekend, Weekday, Summer Weekend, Summer Weekday

Accumulated Precipitation: Full, Winter, Weekday

[5]By our definition, Winter = {January, February, March, October, November, December} and Summer = {April, May, June, July, August, September}.

[6]By our definition, Weekend = {Saturday, Sunday, Monday} and Weekday = {Tuesday, Wednesday, Thursday, Friday}.

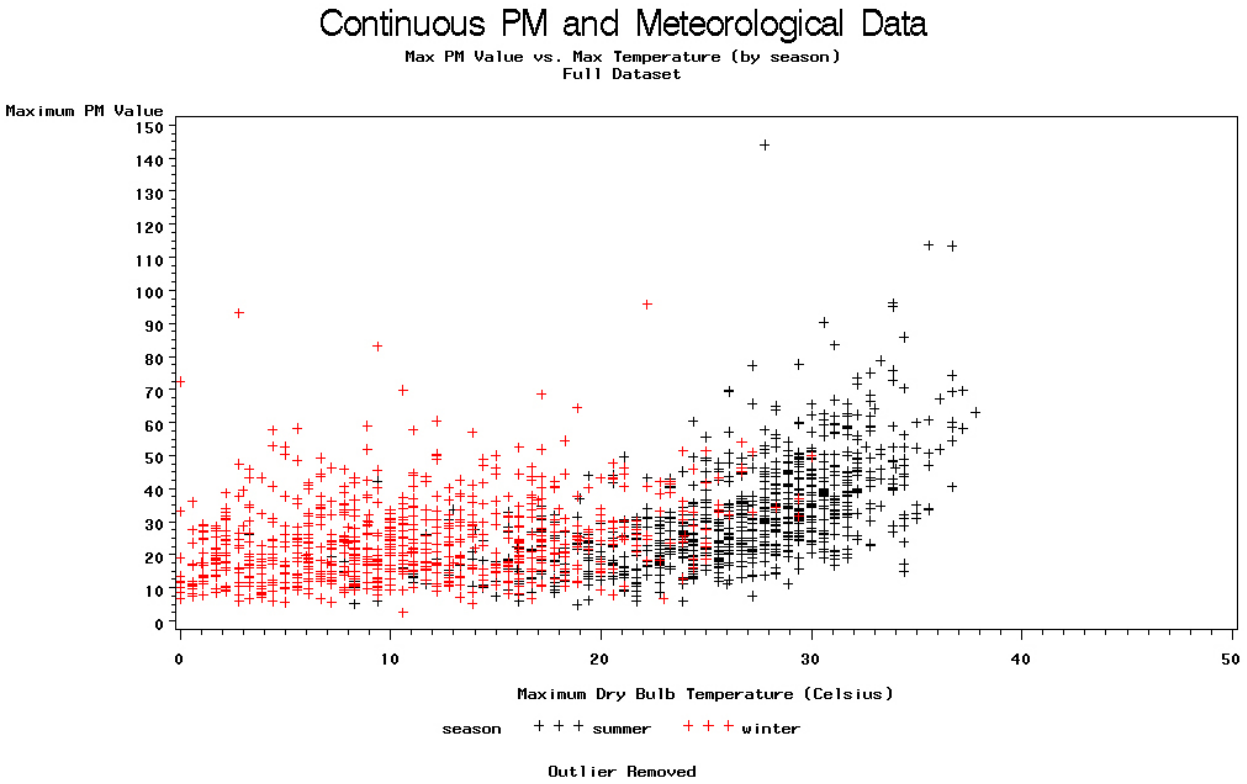

FIGURE 3. Plot of peak temperature vs. peak PM_{cont} values. Notice the random scatter for winter months and noticeable curvature for the summer. This implies the need for stratifications, leading to a piecewise model. Dummy variables would usually be ineffective in such a setting. With dummy variables in a general model, curvature terms would always be considered, even in certain situations where a square term would be insignificant (such as with Peak Temperature, shown in Figure 3). Therefore, stratification and piecewise modeling may capture more information by being able to apply square terms under applicable situations, and simple linear terms with other variables.

Maximum Wind Speed: Full, Summer, Winter, Weekend, Summer Weekend, Summer Weekday, Winter Weekend
Average Wind Speed: ALL
Average Pressure: Summer Weekend
Average Humidity: Full, Winter, Weekend, Weekday, Summer Weekend, Winter Weekend, Winter Weekday
u **vector:** Full, Summer, Winter, Weekday, Summer Weekday, Winter Weekday
v **vector:** ALL
$\mathbf{PM_{yes}}$**:** ALL

Model	R^2 (Abbreviated Dataset)	R^2 (Full Dataset)
Full	.4367	.5523
Summer	.4588	.5623
Winter	.4683	.5476
Weekend	.4213	.5110
Weekday	.4715	.5807
Summer Weekend	.4651	.5134
Summer Weekday	.4751	.5927
Winter Weekend	.4749	.5754
Winter Weekday	.4857	.5404

TABLE 6. Establishment of a direct relationship between stratification complexity and explained variability.

POP: Summer, Summer Weekday

and the follow significant level two interaction terms were found to be useful in prediction with:

(Maximum Wind Speed) × (Average Wind Speed): Winter, Winter Weekend
(POP) × (Average Humidity): Summer, Summer Weekday
(Accumulated Precipitation) × (Average Humidity): Full, Summer, Winter, Weekend, Weekday, Summer Weekend, Winter Weekday

and the following significant level two square terms were found to be useful in prediction with:

Peak Temperature^2: Full, Summer, Weekend, Weekday, Summer Weekend, Summer Weekday
Accumulated Precipitation^2: Full, Summer, Weekend, Summer Weekend
Maximum Wind Speed^2: Weekday, Winter Weekday
Average Wind Speed^2: Full, Summer, Weekend, Weekday, Summer Weekend, Winter Weekday
Average Pressure^2: Summer
Average Humidity^2: Summer, Summer Weekday
v **vector^2:** Full, Winter Weekend, Weekday, Summer Weekend, Winter Weekend, Winter Weekday
PM$^2_{\text{yes}}$: ALL

Some variables were thrown out by backward elimination each time, regardless of stratification. For example, with interaction terms, "(POP) × (Accumulated Precipitation)" and with square terms "(u vector)^2" and "(POP)^2." Perhaps the *greatest trends* can be found with **temperature** driving PM_{cont} values during the **summer** and **humidity**, **wind** and **precipitation** driving PM_{cont} values during the **winter**. Precipitation variables (POP, Average Humidity and Accumulated Precipitation) were somewhat significant, at times, but

Vector Sign	Wind Direction
$+u$, $+v$	$\tan^{-1}(+u/+v)$
$+u$, $-v$	$90 + \tan^{-1}(-v/+u)$
$-u$, $-v$	$180 + \tan^{-1}(-u/-v)$
$-u$, $+v$	$270 + \tan^{-1}(+v/-u)$

TABLE 7. Vectors for analyses are very useful in finding an overall average wind direction. However, wind directions can be easily obtained from the implemented wind vectors.

only with some sort of modifications (transformations, interaction terms, etc.). Temperature, with or without various modifications tended to be very significant in all summer settings. In addition, it can be seen that average wind speed, the v vector (but not the u vector for weekend models, strangely), PM_{yes} and PM^2_{yes}, were very significant variables in all cases, driving much of the regression.

One may find slight bits of oddities, such as a significant u vector in all stratifications, but the v vector was only found to be significant in prediction for models aside from the weekend stratifications. In addition the X_4 interaction term variable containing data of (Accumulated Precipitation) $\times$ (Average Humidity) seemed to have switched in the models. X_4 was seen to be significant for the Summer Weekend model and the Winter Weekday model, but not for Summer Weekday nor Winter Weekend.

Stratifications and the development of a piecewise predictor function have proven to be useful. Additionally, the use of wind components has great explanatory power during winter models, and also in summer models, especially the use of average wind speed instead of peak wind statistics. It is important to observe that with respect to PM_{cont} values, the u vector (East/West) has an inverse relationship and the v vector (North/South) has a direct relationship. This means as wind directions shift from an easterly direction (positive u) to a westerly direction (negative u), PM_{cont} values increase. The same occurs as wind direction shifts from a southerly direction (negative v) to a northerly direction (positive v). Trends with the u and v vector can be observed and further analyses can be made in the future, including oceanic effects (sea-breeze or bay-breeze), and other terrain-generated currents (see table 7).

5. RECOMMENDATIONS

The most important suggestion that can be made would be the presence of co-located meteorological and PM_{cont} sites. The assumption of uniform meteorology is quite a presumption and it would be better and potentially far more accurate if one could rely on co-location. One would expect to find an improvement in R^2 values if such sites were to be implemented and analyzed. Sites, co-located or not, ought to be present throughout the state, especially where there is currently a lack of monitoring sites and data transmission, mostly with the northwest and southeast corners (Calvert County, south Anne Arundel) of the state. This

AQI Index Values	AQI Descriptor	Concentration range (24-hour avg., $\mu g/m^3$)	Color
$0-50$	Good	$0-15.4$	Green
$51-100$	Moderate	$15.5-40.4$	Yellow
$101-150$	Unhealthy for Sensitive Groups	$40.5-65.4$	Orange
$151-200$	Unhealthy	$65.5-150.4$	Red
$201-300$	Very Unhealthy	$150.5-250.4$	Purple

TABLE 8. AQI chart for PM fine. Continuous readings can be placed into categories.

would provide a better understanding of spatial patterns in PM fine. In addition, as PM models are updated, the R^2 values should increase with the prevalence of new, more abundant data. As more data is collected, site-based regression modeling would be very useful, as opposed to a generic model observing general trends as was performed with our study.

Visibility has been proven to be very significant in prediction of PM fine from other similar studies. Such data was available from MDE, but was found to be troublesome in data step production, with varying units and formats. It can be noted that data inputs follow a different system for 2001-04 than with the new system implemented for the 2005-06 data. Connections were very difficult to draw upon, and thus, visibility ultimately was not included for analysis. A revisiting of visibility measurements in future modeling studies may be effective.

The development of software packages[7] will allow the information of various facets of PM fine modeling to be readily available to meteorologists and the interested public. Using this sort of publication, one may generalize PM fine readings into categories, all as defined by an Air Quality Standards Index. This implementation would require the use of ordinal, or in simpler cases, logistic regression models (see Appendix B).

6. Follow-Up Information

Applicable code is available upon request. In addition, follow-up analyses output are also available (Diagnostics output, histograms and boxplots for distributional assumptions, residual plots for heteroscedasticity, and correlation matrices for multicollinearity[8].) PROC REG SAS output is available from modeling allowing us to map daily average PM_{cont} into an "PM_{FRM}-like" value. This can be done with R^2 values of almost 85%. Increases can be made on the explained variation with the inclusion of temperature, precipitation or wind variables. With subsequent models, backwards elimination is used and each step count and

[7]Namely the "CART" system, as implemented by the EPA

[8]The multicollinearity assumption was carefully observed and highly correlated explanatory variables were thrown out. For example, the situation with "Dry Bulb Temperature in Celsius," "Humidity" and "Dew Point" resulted in dew point observations being thrown out since any effect ought to be modeled in a split format, where inference on each individual component can be performed.

R^2 and C_p-mallows[9] Output on all models is available, upon request. Final models are included, however (see Appendix C).

7. Works Consulted

(1) Mendenhall, W., and Sincich, T. 2003. A second course in statistics: regression analysis. 6th ed. Upper Saddle River, NJ: Pearson Education, Inc.
(2) Sutradhar, B. C. 2003. An overview on regression models for discrete longitudinal responses. Statistical Science, 18, 377-393.
(3) Pardo, J. A., Pardo, L., and del Carmen Pardo M. 2005. Testing in logistic regression models based on ϕ-divergences measures. Journal of Statistical Planning and Inference, 136, 982-1006.
(4) Gerig, T. M. "Violations in Assumptions of Regression." NCSU. Harrelson Hall, Raleigh. 16 November 2005.

8. Appendix A: Regression Analysis Theory

We assume

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \tag{8.1}$$

and

$$\beta_i \overset{iid}{\sim} N(\mu_{\beta_i}, \sigma^2_{\beta_i}) \tag{8.2}$$

with x_{ij} collected without error[10]

Usefully, independent variables can be involved with the explanation of variation in a dependent variable. If we consider the use of data vectors, and more generally, data matrices, we can assume a certain multiple regressive linear model formally as

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_k x_k + \epsilon$$

[9]Mallows' C_p is commonly used as a stopping rule in stepwise regression. One potential pitfall of regression is to produce a model that is "overfit," or contains too many independent variables. This can be addressed by observing the explanatory power of all possible models containing subsets of the original group of variables. Instead of defining entrance and exit probabilities, as with backwards elimination (implemented for our modeling, upon request), one would compute regression models using all possible combinations of explanatory variables, observe the C_p Mallows value, defined as: $C_p = \frac{SSE_p}{MSE} - N + 2p$ for p regressors. These values are indicated on the outputs following the full regression model (available upon request). After C_p values are calculated, one can plot these values against the number of variables used, and then to see at what number of variables does the explanatory power level off on the plot. Using this number of variables, one can list R^2 values and pick the model with the largest explained variation.

[10]For x_{ij} collected with error, one may consider measurement error models as a side study. It can be noted that each site operates under its own setup and may differ from site-to-site. Measurement error, if distributional assumptions can be made, may be a prominent worth investigating. For our purposes, and with the support of MDE, measurement error will operate as a relaxed assumption for our study.

or

$$E[y] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{8.3}$$

where we have

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ 1 & x_{13} & x_{23} & \dots & x_{k3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix}$$

$\widehat{\beta}$ can be solved for using elementary matrix algebra, yielding the useful result

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{8.4}$$

Detection diagnostics were performed for this project to detect inaccuracies among regression assumptions. Heteroscedasticity, or non-homogenous variances, can be observed with residual plots of the independent (or explanatory) variables. Detecting model lack of fit with residuals requires plotting of the $\widehat{\epsilon}$ (on the vertical axis) against each of the independent variables, $x_1, x_2, ..., x_k$ on the horizontal axis. Then, plots of $\widehat{\epsilon}$ (on the vertical axis) versus the predicted value $\widehat{y}$ are to be observed. Trends, dramatic changes in variability, or more than 5% of residuals that lie outside $2s$ of 0 should raise concerns. Normal probability plots check the normality (distributional) assumption by plotting residuals against the expected values of the residuals under the assumption of normality. Sorted residuals, $\widehat{\epsilon_i}$ are used to calculate corresponding tail areas given by

$$A = \frac{i - .375}{n + .25}$$

where n is the sample size. Then, the estimated value of of $\widehat{\epsilon_i}$ under normality can be approximated by

$$E(\widehat{\epsilon_i}) \approx \sqrt{MSE}[Z(A)]$$

where MSE is the mean square error for the fitted model and $Z(A)$ is the value of the standard normal distribution (z value) that cuts off an area of A in the lower tail of the distribution. That is,

$$Z \sim N(0, 1) \tag{8.5}$$

and one may recall from standard statistical literature

$$\begin{aligned}\Phi(z) &= F(z) \\ &= P\left[Z \leq z\right] \\ &= P\left[\sqrt{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right) \leq z\right] \\ &= \int_{-\infty}^{z} f(z)du \\ &= \int_{-\infty}^{z} \left(\frac{1}{\sqrt{2\pi}}\right) e^{-z^2/2} du \end{aligned} \tag{8.6}$$

9. Appendix B: Logistic and Ordinal Regression Modeling

We would assume a vector of random variables $\mathbf{Y}$ in a set of interest, under our denoted set of interest Ω are *iid* binomially with (population) parameters n_i and π_i, $\forall i \in \Omega$. That is,

$$Y_i \overset{iid}{\sim} Bin(n_i, \pi_i)$$

It has been shown (proof omitted) that

$$\pi_i = \frac{e^{\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}}{1 + e^{\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}}, \forall i \in \Omega \tag{9.1}$$

Then using our logit function,

$$logit\left(\pi_i\right) = \prod_{i \in \Omega} \left(\begin{array}{c} n_i \\ n_{i1} \end{array} \right) \pi\left(\mathbf{X}_i^T \beta\right)^{n_{i1}} \left(1 - \pi\left(\mathbf{X}_i^T \beta\right)\right)^{n_i - n_{i1}} \tag{9.2}$$

To obtain a good estimator, we can compute the maximum likelihood estimator (MLE), also known as "$\widehat{\beta}$" by computing

$$min\left\{\sum_{j=1}^{2} \sum_{i \in \Omega} \frac{n_{ij}}{N} \log\left(\frac{\frac{n_{ij}}{N}}{\pi_{ij}\frac{n_i}{N}}\right)\right\} \tag{9.3}$$

By using differential calculus, one can observe local (or absolute) maxima over the real line of the "theta-set," defined as

$$\Theta = \{(\beta_0, ..., \beta_k) : \beta_i \in (-\infty, \infty) = \mathbb{R},\ i = 0, 1, ..., k\}$$

Simulation studies using Monte Carlo Markov processes, the bootstrap and the jacknife have all been successfully implemented in the testing and justification said estimators. However, such computations and analyses are beyond the scope of this paper.

10. Appendix C: Final Models

One can implement any level stratification that they wish. It was shown earlier that as stratification complexity increases, explained variability increases. One can employ seasonal stratification, temporal (day of the week) stratification, both stratifications, or simply use the full (unstratified) model. All models use some sort of consolidated daily measurement, not hourly, using the full dataset. So, for the t^{th} day we have:

10.1. **Full Model.**

$$\begin{aligned}\widehat{\overline{PM}}_{\text{cont},t} = &-0.4561 - 0.3343(\text{Peak Temperature}) + 8.1889(\text{Accumulated Precipitation}) - \\ &1.3219(\text{Average Wind Speed}) + 0.4023(v \text{ vector}) + 0.41749(\overline{PM}_{\text{cont},t-1}) + \\ &3.8255\,[(\text{Probability of Precipitation}) \times (\text{Accumulated Precipitation})] - \\ &0.0864\,[(\text{Probability of Precipitation}) \times (\text{Average Humidity})] - \\ &0.1358\,[(\text{Accumulated Precipitation}) \times (\text{Average Humidity})] + \\ &0.0141(\text{Peak Temperature})^2 + 0.0038(\text{Maximum Wind Speed})^2 + \\ &0.0199(\text{Average Wind Speed})^2 + 0.0154(\text{Average Pressure})^2 + \\ &0.0005(\text{Average Humidity})^2 + 0.0121(u \text{ vector})^2 + 5.1283(\text{Probability of Precipitation})^2\end{aligned}$$

10.2. **Seasonal Model.**

$$\widehat{\overline{PM}}_{\text{cont},t} = \begin{cases} 27.3497 - 1.5796(\text{Peak Temperature}) + 7.9379(\text{Accumulated Rain}) - \\ 1.4712(\text{Average Wind Speed}) + 0.4153(v \text{ vector}) + 0.3896(\overline{PM}_{\text{cont},t-1}) + \\ 3.5320\,[(\text{Probability of Precipitation}) \times (\text{Accumulated Precipitation})] - \\ 0.0272\,[(\text{Probability of Precipitation}) \times (\text{Average Humidity})] - \\ 0.1153\,[(\text{Accumulated Precipitation}) \times (\text{Average Humidity})] + \\ 0.0436(\text{Peak Temperature})^2 + 0.0022(\text{Maximum Wind Speed})^2 + \\ 0.0455(\text{Average Wind Speed})^2 & \text{if Summer} \\ \\ -14.9963 - 0.9760(\text{Average Wind Speed}) + 0.8037(\text{Average Pressure}) - \\ 0.1206(u \text{ vector}) + 0.4069(v \text{ vector}) + 0.5815(\overline{PM}_{\text{cont},t-1}) + \\ 7.7747(\text{Probability of Precipitation}) - \\ 0.0262\,[(\text{Maximum Wind Speed}) \times (\text{Average Wind Speed})] + \\ 9.2039\,[(\text{Probability of Precipitation}) \times (\text{Accumulated Precipitation})] - \\ 0.1501\,[(\text{Probability of Precipitation}) \times (\text{Average Humidity})] - \\ 0.1024\,[(\text{Accumulated Precipitation}) \times (\text{Average Humidity})] - \\ 0.0078\,[(u \text{ vector}) \times v \text{ vector})] + 0.0024(\text{Peak Temperature})^2 + \\ 0.0084(\text{Maximum Wind Speed})^2 + 0.0299(\text{Average Wind Speed})^2 + \\ 0.0009(\text{Average Humidity})^2 + 0.0200(u\text{vector})^2 + 0.0352(v \text{ vector})^2 - \\ 0.0072(\overline{PM}_{\text{cont},t-1})^2 & \text{if Winter} \end{cases}$$

10.3. **Temporal Model for Weekly Trends.**

$$\widehat{\overline{PM}}_{\text{cont},t} = \begin{cases} -3.7123 - 0.2844(\text{Peak Temperature}) - 1.4385(\text{Average Wind Speed}) + \\ 0.0560(\text{Average Humidity}) + 0.4118(v \text{ vector}) + 0.3813(\overline{PM}_{\text{cont},t-1}) - \\ 0.1079\,[(\text{Probability of Precipitation}) \times (\text{Average Humidity})] - \\ 0.0197\,[(\text{Accumulated Precipitation}) \times (\text{Average Humidity})] + \\ 0.0125(\text{Peak Temperature})^2 + 0.0057(\text{Maximum Wind Speed})^2 + \\ 0.0375(\text{Average Wind Speed})^2 + 0.0168(\text{Average Pressure})^2 + \\ 0.0200(v \text{ vector})^2 + 9.3215(\text{Probability of Precipitation})^2 & \text{if Weekend} \\ \\ -0.1106 - 0.3564(\text{Peak Temperature}) + \\ 11.6620(\text{Accumulated Precipitation}) - \\ 1.2080(\text{Average Wind Speed}) + 0.4000(v \text{ vector}) + 0.4295(\overline{PM}_{\text{cont},t-1}) + \\ 7.7265\,[(\text{Probability of Precipitation}) \times (\text{Accumulated Precipitation})] - \\ 0.0584\,[(\text{Probability of Precipitation}) \times (\text{Average Humidity})] - \\ 0.1976\,[(\text{Accumulated Precipitation}) \times (\text{Average Humidity})] + \\ 0.0150(\text{Peak Temperature})^2 + 0.0032(\text{Maximum Wind Speed})^2 + \\ 0.0152(\text{Average Pressure})^2 + 0.0006(\text{Average Humidity})^2 + \\ 0.0257(u \text{ vector})^2 + 0.0332(v \text{ vector})^2 & \text{if Weekday} \end{cases}$$

10.4. **Wholly Stratified Model.**

$$
\widehat{\overline{PM}}_{\text{cont},t} = \begin{cases}
2557.0748 - 1.2913(\text{Peak Temperature}) - \\
1.1961(\text{Average Wind Speed}) - 171.7438(\text{Average Pressure}) + \\
0.3232(v \text{ vector}) + 0.3302(\overline{PM}_{\text{cont},t-1}) + \\
0.0238\left[(\text{Maximum Wind Speed}) \times (\text{Average Wind Speed})\right] - \\
0.1010\left[(\text{Probability of Precipitation}) \times (\text{Average Humidity})\right] + \\
0.0385(\text{Peak Temperature})^2 + 2.9084(\text{Average Pressure})^2 + \\
0.0147(v \text{ vector})^2 + 9.4715(\text{Probability of Precipitation})^2 & \text{if Summer, Weekend} \\
\\
31.7084 - 1.7542(\text{Peak Temperature}) - \\
1.8122(\text{Average Wind Speed}) + 0.4907(v \text{ vector}) + \\
0.4109(\overline{PM}_{\text{cont},t-1}) - \\
0.0270\left[(\text{Probability of Precipitation}) \times (\text{Average Humidity})\right] + \\
0.0465(\text{Peak Temperature})^2 + 0.0647(\text{Average Wind Speed})^2 & \text{if Summer, Weekday} \\
\\
-19.8939 - 11.1972(\text{Accumulated Precipitation}) - \\
1.1251(\text{Average Wind Speed}) + 0.9526(\text{Average Pressure}) + \\
0.4174(v \text{ vector}) + 0.5825(\overline{PM}_{\text{cont},t-1}) + \\
12.3872(\text{Probability of Precipitation}) - \\
-0.0524\left[(\text{Maximum Wind Speed}) \times (\text{Average Wind Speed})\right] + \\
9.6897\left[(\text{Probability of Precipitation}) \times (\text{Accumulated Precipitation})\right] - \\
0.1748\left[(\text{Probability of Precipitation}) \times (\text{Average Humidity})\right] + \\
0.0016(\text{Peak Temperature})^2 + 0.0153(\text{Maximum Wind Speed})^2 \\
0.0784(\text{Average Wind Speed})^2 + 0.0008(\text{Average Humidity})^2 \\
0.0227(v \text{ vector})^2 - 0.0069(\overline{PM}_{\text{cont},t-1})^2 & \text{if Winter, Weekend} \\
\\
0.0739 + 0.0855(\text{Maximum Wind Speed}) - \\
1.1008(\text{Average Wind Speed}) - 0.1400(u \text{ vector}) + \\
0.3809(v \text{ vector}) + 0.6023(\overline{PM}_{\text{cont},t-1}) + \\
7.5079\left[(\text{Probability of Precipitation}) \times (\text{Accumulated Precipitation})\right] - \\
0.0739\left[(\text{Probability of Precipitation}) \times (\text{Average Humidity})\right] - \\
0.0844\left[(\text{Accumulated Precipitation}) \times (\text{Average Humidity})\right] + \\
0.0032(\text{Peak Temperature})^2 + 0.0108(\text{Average Pressure})^2 + \\
0.0008(\text{Average Humidity})^2 + 0.0299(u \text{ vector})^2 + \\
0.0447(v \text{ vector})^2 - 0.0076(\overline{PM}_{\text{cont},t-1})^2 & \text{if Winter, Weekday}
\end{cases}
$$

Department of Marine, Earth and Atmospheric Sciences and Department of Statistics,
North Carolina State University, Raleigh, NC.