OPTIMIZING THE WIND VARIABLE FORMULATION IN A MULTIPLE LINEAR REGRESSION MODEL OF PARTICLE NUMBER CONCENTRATION

by

Caitlin Collins

Senior Honors Thesis

Advisor: John Durant

Department of Civil and Environmental Engineering

Tufts University

Medford, Massachusetts

April 26, 2012

ABSTRACT

Airborne ultrafine particles (UFP; diameter < 100 nm) are typically present at high levels near highways. Due to their small size, UFP may have more adverse health impacts than larger particles. The concentration of UFP can vary widely over a small area, especially in an urban setting, so many monitoring locations are necessary to accurately estimate UFP concentrations on an urban scale. Because of this, a model of particle number concentration (PNC; a proxy for UFP) based on mobile monitoring data can present a less costly alternative for estimating UFP levels than deploying a monitoring network. Wind is known to have an effect on PNC, so it is important that any model of PNC with high temporal resolution include an effective wind variable. Because wind is a vector quantity (has both direction and magnitude), it can be difficult to represent in a model. Many air pollution models entirely omit the effects of wind or only include wind speed as a variable. The objectives of this thesis were (1) investigate a suite of different wind variables used in air pollution modeling by including them in multiple linear regressions of PNC, (2) compare the regressions using goodness-of-fit statistics such as adjusted R^2, and (3) validate the regressions by using regression diagnostics to determine whether modeling assumptions were met. To investigate the relative merits of different wind variables, a multiple linear regression of the logarithm of hourly PNC (Adj. $R^2 = 0.22$) was augmented with one of four types of wind variables: wind speed, wind sectors, a vector-based variable, or a Fourier-based variable. Wind sectors based on wind direction relative to the highway performed best (Adj. $R^2 > 0.30$), while the vector-based variables performed poorly (Adj. $R^2 < 0.25$; less than wind speed alone). The wind sectors may have performed better than vector-based variables because only one weather station was used, which meant that local turbulence was not accounted for, thus leading to incorrect wind direction input to the vector variables. The highway-relative wind sectors may have performed better than the conventional sectors because they took the effect of the highway, which is considered to be a major source of UFP within the study area, into account. The findings from this research will inform a regression model of PNC that will aid in assessing exposure to UFP for the study participants of CAFEH.

TABLE OF CONTENTS

FIGURES iv

TABLES v

INITIALISMS vi

1.0 INTRODUCTION 1

1.1 Health Effects of Traffic Pollution 1

1.2 The CAFEH Study 1

1.3 Use of Wind Variables in Air Pollution Models 2

1.4 Multiple Linear Regression 4

1.5 Problem and Objectives 7

2.0 MATERIALS AND METHODS 8

2.1 Study Area and Data Collection 8

2.2 Data Compilation and Averaging Procedures 10

2.3 Pre-Regression Analysis 11

2.4 Development of the Regression Datasets 13

2.4.1 Vector variables 14

2.4.2 Wind sectors 14

2.4.3 First-order Fourier approximation 16

2.5 Regression 16

3.0 RESULTS AND DISCUSSION 18

3.1 Pre-regression Data Analysis 18

3.2 Regression Results 25

3.2.1 Effect of the 1-hour-lagged PNC term 26

3.2.2 Performance of vector variables 30

3.2.3 Performance of wind sectors 30

3.2.4 Performance of first-order Fourier approximation 30

4.0 CONCLUSIONS 32

4.1 Optimal Wind Variable 32

4.2 Autocorrelation of the Data 33

5.0 ACKNOWLEDGMENTS 34

REFERENCES

APPENDICES

FIGURES

1. Example map of PNC measured by the TAPL 2
2. Map of the study area and monitoring route in Somerville, MA 8
3. Averaging method 11
4. Application of the dot product to wind variables 14
5. Depictions of traditional and highway-relative wind sectors 15
6. Translated sine function to illustrate Fourier variables 16
7. Normal probability plot of the logs of averaged PNC 18
8. Traffic representativeness 19
9. Wind speed and direction representativeness 20
10. Seasonal PNC variation 21
11. PNC variation with wind 22
12. Autocorrelation function for MAC fixed site; lags up to 24 hours 24
13. Autocorrelation function for MAC fixed site; lags up to 240 hours 24
14. Residuals vs. predicted value 26

TABLES

1. Variable descriptions 13

2. Regression results for models run without 1-hour-lagged PNC term 28

3. Regression results for models run with 1-hour-lagged PNC term 29

INITIALISMS

ACF	Autocorrelation Function
CAFEH	Community Assessment of Freeway Exposure and Health
GPS	U.S. Global Positioning System
I-93	Interstate 93
LUR	Land Use Regression
MAC	Mystic Activity Center
NO_2	Nitrogen Dioxide
PNC	Particle Number Concentration
TAPL	Tufts Air Pollution Laboratory
UFP	Ultrafine Particles
VIF	Variance Inflation Factor

1.0 INTRODUCTION

1.1 Health Effects of Traffic Pollution

Exposure to traffic-related air pollution is associated with elevated risks of heart and lung diseases (Brunekreef et al., 1997; Brugge et al., 2007; Gauderman et al., 2007; Hwang et al., 2005; McConnell et al. 2006; Van Vliet et al., 1997). One group of air pollutants that may contribute to these risks is ultrafine particles (UFP; diameter <100 nm), which are typically present at elevated levels near highways and busy roadways (Durant et al., 2010; Zhu et al., 2002), more toxic than larger particles (Oberdörster et al., 2005; Sioutas et al., 2005), and able to penetrate from the lungs into the blood system (Oberdörster et al., 2005). To date, no studies have determined whether exposure to near-highway UFP is associated with human health impacts (Brugge et al., 2007).

1.2 The CAFEH Study

The Community Assessment of Freeway Exposure and Health (CAFEH) aims to quantify the association between cardiovascular health and exposure to near-highway air pollutants. Findings from CAFEH could contribute to better understanding of health problems in people living near highways, and could also help inform future policies regarding traffic pollution.

As part of the CAFEH study, approximately 200 people in Somerville, Massachusetts, completed health surveys and gave blood samples for analysis of markers of cardiovascular health such as C-reactive protein and fibrinogen. In addition, particle number concentration (PNC) data was collected via mobile monitoring in the study area by the Tufts Air Pollution Monitoring Laboratory (TAPL) to help assess exposure of each of the participants to near-highway UFP, for which PNC is a proxy. The mobile lab allows data collection over a large area but, because the mobile lab moves continuously through the

study area, the resulting dataset lacks detailed information on temporal variation at fixed locations. Thus, a multiple linear regression model will be constructed from the data which will ultimately be used to assign exposures of participants to PNC on both a time- and space-resolved basis. Figure 1 is an example of data taken over approximately one hour in the Somerville neighborhood with the Tufts Air Pollution Monitoring Laboratory (TAPL). Although the measurements are only temporally dense over short periods (4-6 hours per monitoring day), the measurements are very spatially dense. This is an advantage of mobile monitoring (Hagler et al., 2010) and should help make modeling of PNC in the study area more effective (Briggs, 2005; Jerrett et al., 2005).

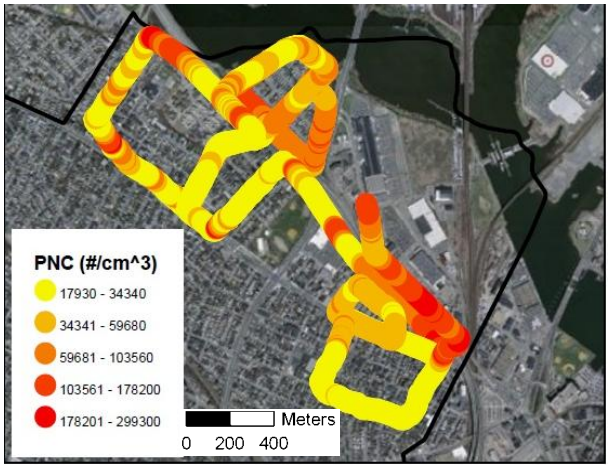

Figure 1: Map of PNC measured by the TAPL from 5:50 to 6:41 AM on January 6, 2010, wind from WNW at 4.1 m/s. Note the elevated levels (red color) near the highway. (Figure prepared by Luz T Padró-Martínez. Orthophotos from MassGIS.).

1.3 Use of Wind Variables in Air Pollution Models

The most common modeling techniques used to assess air pollution exposure in an urban environment are spatial interpolation, line dispersion models, and land use regressions (LUR). Spatial interpolation uses geostatistical methods to estimate concentrations of a pollutant between multiple measurement

locations. While this method can provide a reasonable estimate of exposure, it requires a dense network of monitoring sites to be effective, rather than relying on the relationships between air pollution and possible predictors such as meteorology and distance to source (Jerrett et al., 2005). Dispersion models base their predictions on theoretical equations such as the Gaussian plume model, and as such rely heavily on meteorological inputs. Error in the meteorological inputs for a dispersion model may introduce uncertainty into the model and negatively affect its performance (Seaman, 2000). LUR is a regression technique that models the spatial distribution of the pollutant of interest using spatial predictors such as land use and land cover, traffic density, or distance to a power plant. LUR models typically do not take meteorological factors such as wind speed or direction into account (Hoek et al., 2008).

Some health studies, such as birth cohort studies, may require a temporal resolution finer than a year (Hoek et al., 2008). One way to achieve finer temporal resolution is to combine LUR and dispersion model approaches into a hybrid model that includes both spatial and temporal factors. For hybrid models of PNC, the inclusion of wind will be important, since PNC is known to be affected by both wind speed (Zhu et al., 2002) and wind direction (Klems et al., 2010). Wind speed affect dispersion and mixing within an area and alter the travel time of pollutants from one place to another. Wind direction can change the effect of local air pollution sources, since an area that is downwind of a source is likely to experience much higher concentrations than an area upwind of a source.

A few hybrid models have incorporated wind as an explanatory variable. Arain et al. (2007) improved on a traditional LUR model by using the dot product to incorporate highway-relative wind speed and direction into their model of nitrogen dioxide (NO_2). Su et al. (2008) predicted nitric oxide and NO_2 using a modified box model with wedge-shaped regions whose size was determined by meteorological factors such as wind speed and Pasquill stability class. They found that the model was only more effective than traditional LUR when hourly pollution and weather data from 19 nearby

monitoring stations was used to train the model, rather than annual average pollutant measurements from over 100 densely distributed passive samplers. Zwack, et al. (2011) used mobile monitoring data to develop a generalized additive model for UFP which included wind speed but not wind direction as an explanatory variable.

Based on the results of recent modeling studies and the well-known physical association between wind and air pollution, adding wind-based variables to hybrid regression models may allow for increased model prediction power.

1.4 Multiple Linear Regression

Multiple linear regression is a model that attempts to predict some quantity, termed the "dependent variable," as a linear combination of several related predictors, termed "explanatory variables." A multiple linear regression model may be expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots + \beta_k x_k + \varepsilon$$

where y is the dependent variable, β_0 is the intercept, the x_i's are explanatory variables, the β_i's are coefficients for each variable ("model parameters"), and ε is the unexplained variation in y (also called the "error"). Linear regression makes several assumptions about the data and the model error:

1. y is linearly related to each x_i when all other variables are held constant
2. The coefficients are fit using data that is representative of the data the model will be predicting
3. The residual error has constant variance (is homoscedastic) and does not depend on any model variable, time, space, or the model predictions
4. The residuals (error at each predicted point) are independent of each other in space and/or time
5. The residuals are normally distributed
6. The explanatory variables are not correlated with each other

Fulfilling assumptions 1 – 4 and 6 is important if one is trying to determine the best model of y given x_1, x_2, … x_k. Assumption 5 is also very important, but is only necessary for hypothesis tests (e.g. t-test, F-test) or estimating confidence intervals (Helsel and Hirsch, 2002, pp. 222-25).

Assumption 1 is the most important and must be met for the model to be meaningful. It can be addressed by using variable transformations to linearize relationships between PNC and other variables, or by switching to a different model form. If assumption 2 is violated, the usefulness of the model to predict PNC at times that were not well-represented in the mobile monitoring data will be limited.

Fulfilling assumptions 3-5 is important for accurate estimation of statistical measures of the regression, such as confidence intervals, variances, and hypothesis tests.

A model's residual error is homoscedastic if it is unvaryingly random throughout the model. If the model error's mean or variance has a trend (is heteroscedastic), it could affect the standard error of model outputs and the estimated confidence intervals for the model might not be accurate. If the residuals are homoscedastic, they will not show a trend in mean or variance when they are plotted against model predictions, variables, or time. Heteroscedasticity can either be inherent in a dataset or can reflect problems with model form such as nonlinearity or correlation of model variables.

A model can violate assumption 4 by having spatial or temporal autocorrelation. Spatial autocorrelation is the correlation of residuals with other nearby residuals. Air pollution data is known to be highly autocorrelated spatially, and many LUR models have assessed spatial autocorrelation using Moran's I (Hoek et al., 2008; Su et al., 2008; Arain et al., 2007). Temporal autocorrelation, or the correlation of residuals with their time-lagged selves, is a common problem with time-series air pollution models (Mølgaard et al., 2011) and can affect measures of variance and hypothesis tests of parameter significance such as the t-test or F-test. Temporal autocorrelation of a variable can be assessed by calculating the sample autocorrelation function (ACF) for increasing length of lag time. The ACF for a lag of size *n* is equal to the correlation between values of a variable and values of the variable *n* lags earlier. By plotting the ACF versus lag size, a correlogram plot is produced which shows the autoregressive structure of the variable.

Some autocorrelation can be corrected for by adjusting variables for seasonality or a time trend (Zwack et al., 2011), and autocorrelation at a lag of size *n* can be corrected for by adding *n*-lagged values of the dependent variable to the model as was done by Clifford et al. (2011). Autocorrelation can also be corrected for by modeling the difference between the dependent variable and pollutant concentrations at a nearby stationary site, as was done in Hoek et al. (2011). Fruin et al. (2008) reduced autocorrelation in their model of on-highway UFP via an intelligent averaging scheme. If temporal autocorrelation is a great concern, a different model form can be used that allows for autocorrelated errors (Mølgaard et al., 2011). The issue of temporal autocorrelation is more frequently addressed for regression models with a time resolution of hours or days, so many LUR models do not take it into account.

In order for confidence intervals and significance tests to be meaningful for a model, the error must be normally distributed. If the error is highly skewed, it can reflect a serious problem with the model, such as non-normality of the dependent variable or nonlinearity of the model. Non-normality of the error can also be caused by data points with high leverage. Normality of the residuals can be examined using probability plots and hypothesis tests.

The interrelatedness of model variables, also known as multicollinearity, can cause serious problems with parameter estimation. Variable coefficients may change greatly with the addition or deletion of a variable or data point, or may have the wrong sign. Multicollinearity can be detected by looking at the variance inflation factor (VIF) for each variable, which is related to the R^2 of the regression of that variable against all of the other variables. The VIF for every variable should be near to one, and a VIF of greater than ten indicates serious multicollinearity (Helsel & Hirsch, p.305-6, 2002).

1.5 Problem and Objectives

Many air pollution regression models focus on predicting annual average concentrations, which may not meet the needs of all epidemiological studies and may fail to accurately represent the effect of more finely-resolved temporal effects of meteorology on air pollution levels. When data is available at a finer temporal resolution, models based on such data may allow for increased accuracy in estimating exposures. One way to improve an air pollution regression model with high temporal resolution is to add variables that account for wind effects. My goal is to determine the optimal way to include a wind variable in a multiple linear regression model for PNC based on mobile monitoring data and a single weather station. The objectives are to (1) investigate a suite of different wind variables used in air pollution modeling by including them in multiple linear regressions of PNC, (2) compare the regressions using goodness-of-fit statistics such as adjusted R^2, and (3) validate the regressions by using regression diagnostics to determine whether modeling assumptions were met. Ultimately, my findings will help determine the formulation of the wind variable in the final CAFEH Somerville regression model.

2.0 MATERIALS AND METHODS

2.1 Study Area and Data Collection

The data herein is from mobile monitoring performed near Interstate 93 (I-93) in Somerville, Massachusetts (Figure 2). Monitoring was performed on both sides of I-93, which is 40 m wide with 4 lanes in each direction and is elevated 5 m above street level through much of the study area. There is a 3-m-high concrete noise barrier on the eastern side of the highway, which may have some impact on pollutant transport. Broadway, an arterial, carries two bus lines through the study area.

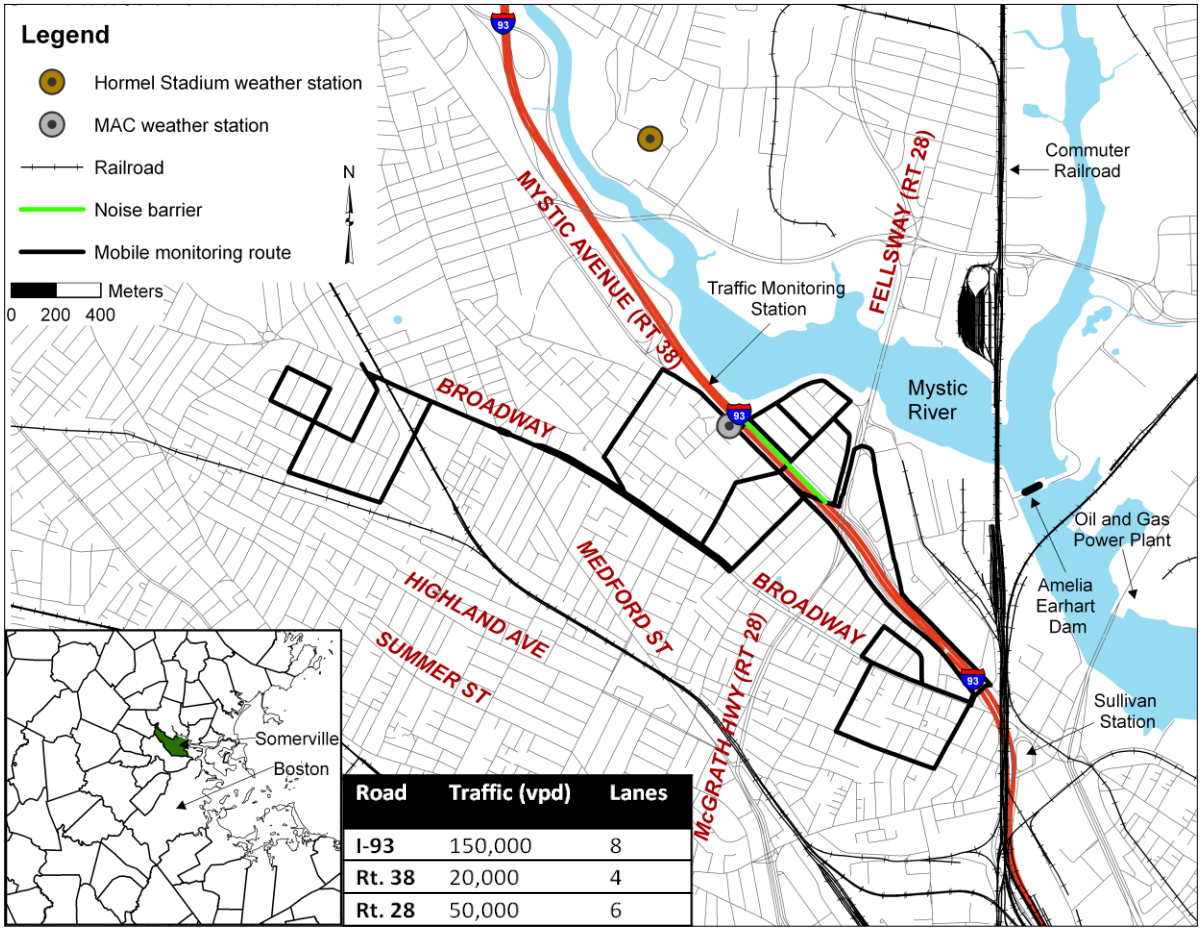

Figure 2: Map of the study area and monitoring route in Somerville, MA. Major streets and other local air pollution sources are labeled. (Padró-Martínez et al., 2012)

Two other major roads through the area are Massachusetts Route 38, which runs along the west side of I-93 through the study area and Massachusetts Route 28, which crosses underneath I-93 (see Figure 2). Other local sources of air pollution may include residential heating, the Amelia Earhart Dam, the bus depot at Sullivan Station, the commuter rail line and an oil and gas-fired power plant.

The TAPL monitored air quality on a fixed route (Figure 2), which the TAPL traversed at 16-32 kph as rapid-response instruments measured a suite of gaseous and particle-phase pollutants. PNC (>4 nm; CPC 3775, TSI, Shoreview, MN) with a time resolution of 1 second was measured as a surrogate for UFP, and measurements were matched to GPS location at 1-second intervals. The route included both a near-highway component as well as urban background area more than 1000 m from the highway. Monitoring was conducted over a range of meteorological and traffic conditions that spanned all four seasons at different times of day, although monitoring during certain conditions (snow, heavy rain, late-night/early-morning, and weekends) was less frequent due to safety concerns or logistical issues. A total of 55 mobile monitoring trips were made through the neighborhood in 2009-2010. Additional stationary monitoring was done with the TAPL several times over the course of monitoring by parking it near the Blessing of the Bay boathouse (near the MAC in Figure 2) for a 24-hour interval and running the instruments with electricity from a nearby building.

Meteorological data were recorded every five minutes at the Hormel Stadium weather station (height ~35 m, wunderground.com) in Medford, which is ~1.5 km from the center of the study area. Data from a meteorology and air pollution (PNC, particle-bound PAH) station at the Mystic Activity Center (MAC), about 100 m from I-93, was used when data from Hormel was not available. The MAC site was not used for weather data in general because it was not operating for all days and may have been influenced by the MAC building or the nearby raised highway.

2.2 Data Compilation and Averaging Procedures

To merge the GPS and PNC data by timestamp, data was added to a MySQL (www.mysql.com) database using a PHP script. Data was then subjected to a quality control procedure in Excel using VBA macros and time series plots. Quality control included the following steps:

(1) removing data collected during instrument malfunctions noted in a daily monitoring log,

(2) correcting timestamps to account for measurement lag of the CPC, and

(3) removing periods of possible self-monitoring.

Data was removed due to self-monitoring if the wind was faster than the TAPL speed and blowing from behind the TAPL, or if the speed of the TAPL was less than 3 mph (i.e. at intersections and in traffic). Overall, less than 10% of data from each monitoring day was removed to avoid inclusion of data that may have reflected self-monitoring. Pollutant spikes due to nearby vehicles in traffic were not removed.

Pollutant concentrations were mapped to their GPS-measured location using ArcGIS 9.3.1 and data layers from MassGIS (http://www.mass .gov/mgis/massgis.htm). The data points were then assigned to the nearest road centerline and each road centerline was split into 20-m segments. Spatial variables (e.g. datapoint location, distance to highway) were then calculated from the centroid of each 20m segment (Figure 3). This helped to remove short-term variation and create the effect of fixed sites, which are computationally easier to deal with since their location does not change over time. The averages were then log-transformed, averaged by hour, and merged with other explanatory variables in SAS to obtain the final regression dataset.

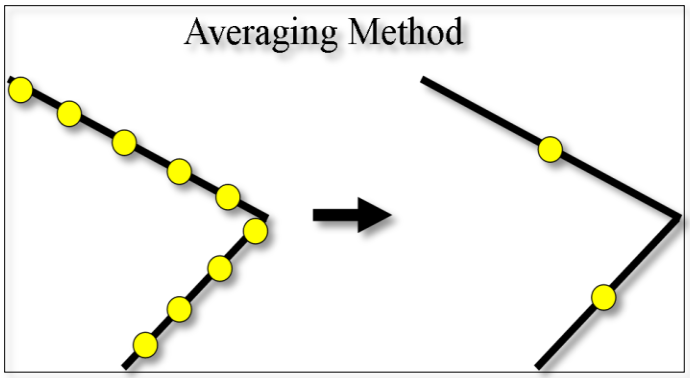

Figure 3: Points along each 20-m road segment were averaged together and the averaged value was assigned the coordinates of the segment centroid. (Figure by Allison Patton)

2.3 Pre-Regression Analysis

Because monitoring was conducted over relatively short periods of time rather than continuously, the data collected was likely not representative of the whole year in terms meteorology. Prior to regression, the distributions of the values of measured variables were inspected to determine whether the dataset was representative and identify conditions that may have been underrepresented in the data. Data was also grouped by distance to the highway using ArcGIS and box plots of the grouped data were made (Padró-Martínez et al., 2012). These box plots were used to visualize variation of PNC over several broad categories by separating box plots for different conditions.

Since environmental data, including PNC, is usually highly skewed it was determined that it might be necessary to transform the dependent variable to meet assumptions of normality and linearity. A natural log transform was used since PNC is usually assumed to be lognormally distributed (Hinds, p.90-5, 1999). MINITAB was used to conduct a hypothesis test on the natural log of PNC for normality (α=0.05) based on the Kolmogorov-Smirnov statistic (in MINITAB: Stat->Basic Statistics->Normality Test).

Assessing autocorrelation in the mobile monitoring data was difficult due to the spatio-temporal distribution of the dataset. Spatial autocorrelation can only be measured when a dataset contains measurements in multiple places at one time, which is not possible with a single mobile monitoring

vehicle. Temporal autocorrelation, on the other hand, can only be fully characterized when the data are a regular continuous time series at one location as in Mølgaard et al. (2011). Data collected during mobile monitoring vary over space and time simultaneously and are a highly irregular time series due to the intermittent nature of the data collection. Because of this, temporal autocorrelation was assessed using more oblique methods than are used in time-series analysis.

Autocorrelation of the errors of a model is often a product of autocorrelation of the dependent variable of the model. By assessing the temporal autocorrelation present in the measured values for PNC, an idea could be formed of the potential for autocorrelation in the residuals of the final model. To obtain a general idea of the nature and magnitude of temporal autocorrelation of PNC in the study area, the temporal autocorrelation of the PNC data from the fixed site (MAC) was calculated for hourly lags up to 24 hours and up to 240 hours using MINITAB (in MINITAB: Stat->Time Series->Autocorrelation). Next, a basic autocorrelation test of the mobile monitoring data was conducted by calculating the 1-hour-lagged PNC for road segments and times when there was PNC data at a single road segment for two or more hours in a row, and then regressing the lagged PNC data with the current data (see SAS code in Appendix 1). Autocorrelation of greater lags was not investigated in the mobile monitoring dataset due to practical reasons: the TAPL rarely visited the same location each hour for more than two hours in a row, so the dataset would have been significantly smaller for each lag that was calculated. Because of this, the MAC autocorrelation calculations were used to characterize higher-order temporal autocorrelation within the study area.

Variable Descriptions	
Original Variables	**Description**
wsmps	wind speed in meters per second
TempF	temperature in degrees Fahrenheit
disti93	distance from the edge of I-93, meters
Hwy_Direction	Direction from the point to the highway, degrees clockwise from North
wdir	wind direction, degrees clockwise from North
lag_logPNC_Mean	1-hour-lagged PNC at the same road segment as the current data point
Vector Variables	**Description**
vdp	dot product of unit vectors of wind speed distance to I-93
vwsd93dp	wsmps*vdp/disti93
vd93dp	disti93/(vdp+1.1)
vws	wsmps*vdp
vwssq	wsmps*wsmps*vdp
vnvws	(1/wsmps)*vdp
Fourier Variables	**Description**
cosdir	cosine of wind direction in radians
sindir	sine of wind direction in radians
Boolean Variables	**Description**
onmaj	0=not on major road, 1=on major road
Downwind	1=downwind conditions, 0=not downwind conditions
Upwind	1=upwind conditions, 0=not upwind conditions
ParallelNW	1=wind parallel to I-93 from the northwest, 0= not parallel from northwest
ParallelSE	1=wind parallel to I-93 from the southeast, 0= not parallel from southeast
Sector1	1=wind from the northeast quadrant, 0= not from NE quadrant
Sector2	1=wind from the northwest quadrant, 0= not from NW quadrant
Sector3	1=wind from the southwest quadrant, 0= not from SW quadrant
Sector 4	1=wind from the southeast quadrant, 0= not from SE quadrant

Table 1: Descriptions of original and derived variables used in regressions. For SAS code definitions of variables, see Appendix 2.

2.4 Development of the Regression Datasets

The data used for regression were a combination of original data and derived variables (Table 1). The original data included meteorological data such as wind speed and temperature in degrees Fahrenheit, as well as spatial variables such as distance and direction to the highway. Derived variables included vector variables, wind sectors, and Fourier-based variables. For SAS code definitions of all wind variables, see Appendix 2.

2.4.1 Vector variables

The scalar product, or dot product, of two vectors allows calculation of the component of one vector in the direction of the other. By taking the dot product of a wind vector, $\vec{w}$, and the unit vector pointing from a pollution receptor point **x** to the highway, $\vec{h}$, we can determine the component of the wind vector that is traveling perpendicular to and away from the highway towards **x** (Figure 4), which can provide an estimate of the distance a particle travels from the highway to a location. A number of variables were derived from this relationship, including the basic vector variable *vdp*, which is the dot product of the unit vector of the wind direction and the unit vector pointing from a segment to I-93.

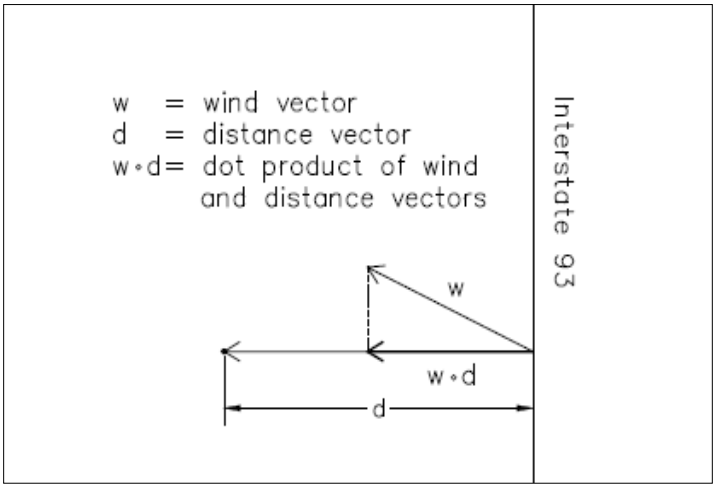

Figure 4: The basic wind vector variable, vdp, is the component of the wind vector that is pointing directly from the highway to the data point.

2.4.2 Wind sectors

Categorical variables can be used to characterize differences in the mean value of PNC measurements for different wind directions. This allows the effect of various sources to be accounted for. For example, PNC measurements downwind of the highway or some other source of UFP will generally be higher than measurements when the wind was blowing in the other direction. Adding a variable that is equal to 1 for certain wind conditions and 0 for all other conditions modifies the intercept for that condition only.

For this investigation, two types of sectors were used: conventional wind sectors and highway-relative wind sectors.

Conventional wind sectors partition the dataset based solely on wind direction. In this study, the wind rose was split into four quadrants, with each sector corresponding to wind out of that quadrant (Figure 5). For the highway-relative sectors, wind direction relative to the highway was used. Data was categorized as either downwind of I-93, upwind of I-93, wind parallel to I-93 from the northwest, or wind parallel to I-93 from the southeast (Figure 5).

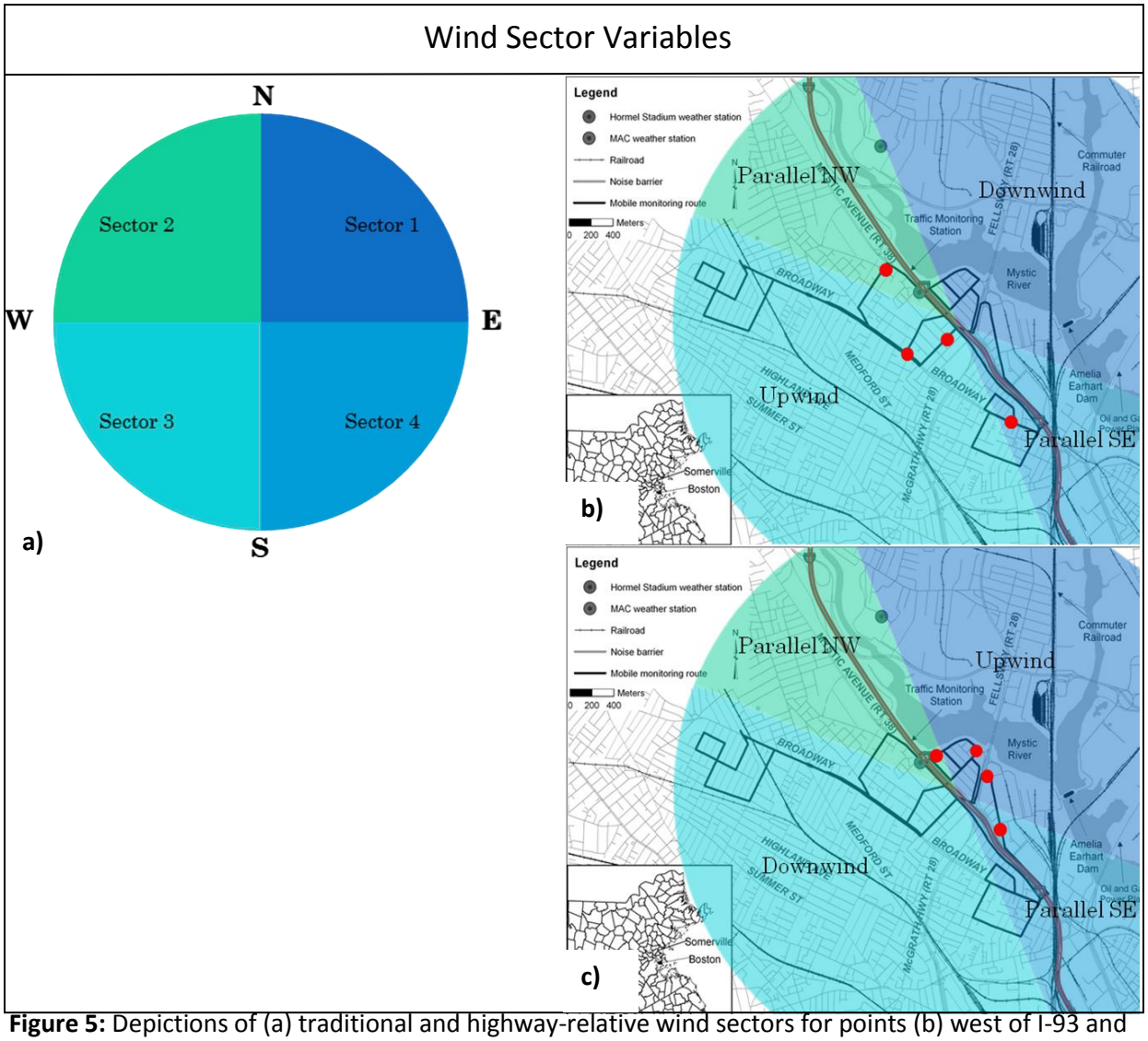

Figure 5: Depictions of (a) traditional and highway-relative wind sectors for points (b) west of I-93 and (c) east of I-93 (hypothetical datapoint positions shown as red dots). Upwind/downwind conditions are reversed for points on opposite sides of I-93; for example, points west of I-93 are downwind when the wind is out of the east, while points east of I-93 are downwind when the wind is out of the west.

2.4.3 First-order Fourier approximation

We would expect pollution levels to vary according to wind direction due to the presence of several possible sources of UFP in the study area, with maximum PNC levels when the wind is blowing directly to a location from a local source of UFP. Since wind direction is represented periodically using the 360° compass, any response to it is also periodic and thus can be approximated by a Fourier series. In this study, a first-order Fourier approximation was used and was accomplished by including the cosine and sine of wind direction, which would allow the regression to pick model coefficients for each one, effectively approximating a translated sine function. Figure 6 illustrates the theoretical relationship between a translated sine function and the sum of a sine and cosine.

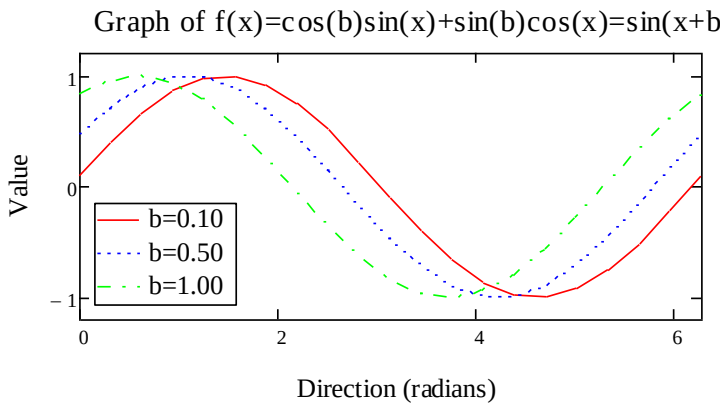

Figure 6: Translated sine function. The maximum of the function moves closer to zero (north, in the case of wind direction) as b becomes larger.

Using this approach assumes that there is one wind direction that is associated with maximum PNC (i.e. a "source"), and that the minimum PNC is found when the wind blows in the opposite direction.

2.5 Regression

Although the focus of this investigation was on wind variables, a set of extra variables was included in every multivariate regression to ensure that the model explained the data with an adjusted R-squared of at least 0.2. The models considered in this investigation were of the form

$$logPNC\_Mean = \beta_0 + \beta_1 \cdot TempF + \beta_2 \cdot onmaj + \beta_3 \cdot lag\_logPNC\_Mean + \sum_{i=4}^{n} \beta_i \cdot [wind\ variable]_i + \varepsilon$$

Where n=4, except in the case of wind sectors. The dependent variable is the spatially- and temporally-averaged logs of PNC (logPNC_Mean), the β_i are model coefficients for each explanatory variable, which were determined using multivariate regression in SAS, and ε is the error.

Linear regression was based on six basic assumptions about the data and the residual error ε (see Section 1.4). The validity of these assumptions was tested using plots of the residuals and regression diagnostic statistics.

Multivariate regressions were run in SAS using PROC REG and the predictions and studentized residuals were output to a CSV dataset. The datasets were imported into MINITAB, where the residuals were examined for homoscedasticity and their normality was tested. Residuals were not tested for spatial or temporal autocorrelation for the same reasons described in Section 2.3, since the residual datasets had the same simultaneous spatio-temporal variation as the mobile monitoring dataset.

To test for homoscedasticity, the residuals were plotted against each explanatory variable, time in hours since midnight on January 1, 2009, and road segment ID (a measure of location). To test for normality, MINITAB was used to conduct a hypothesis test (α=0.05) based on the Kolmogorov-Smirnov statistic (in MINITAB: Stat->Basic Statistics->Normality Test). Because this study is mostly interested in relative performance of models rather than in the predictive power of any model, the normality of the residuals after PNC was transformed was given only a cursory examination to check for extreme skewness.

VIFs were inspected for a few sample regressions to check for inter-variable relationships. Multicollinearity was not expected to be a serious concern with this study, since the variables used in any single regression had almost no theoretical relationship to each other, and thus no expected correlation.

3.0 RESULTS AND DISCUSSION

3.1 Pre-regression Data Analysis

Transforming the dependent variable is one way to ensure linearity of the model and increase normality of the residuals. PNC is commonly transformed via logarithm. A normal probability plot of the averaged natural log of PNC data was generated in MINITAB (Figure 7), and a Kolmogorov-Smirnov (KS) hypothesis test (α=0.05) showed that the data were not normally distributed. Despite this, it was determined that the log transformation would most effectively reduce the skew of the PNC data while still being easy to implement.

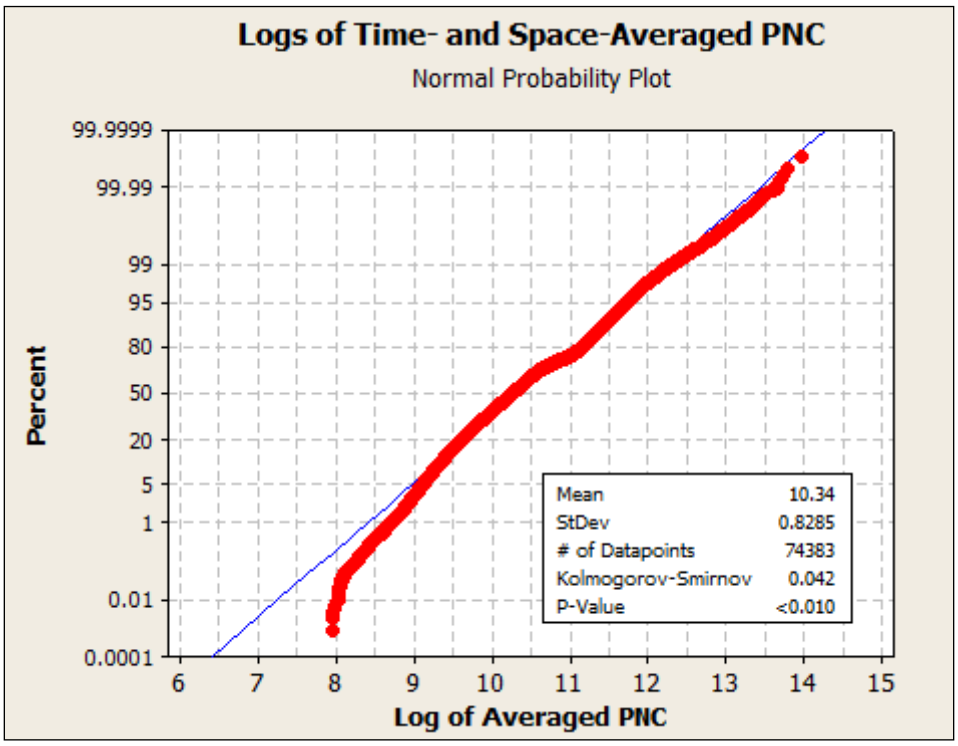

Figure 7: Normal plot of the natural log of PNC. Although the left tail of the distribution does depart from normality, the data nevertheless looks linear enough on the plot to justify a log transformation for the sake of simplicity.

Graphs were used to compare conditions during the monitoring hours to conditions for the entire year (Figures 8-9). While the comparisons of meteorology over the study period vs. the entire year did

not show that the two periods were very similar, the periods were similar enough to support further analysis. Some low-traffic periods (weekends and late night/early morning) were underrepresented or ignored completely in the monitoring schedule, which may cause the model to have little or no prediction power for those time periods. Since the purpose of this investigation was to compare the relative performance of several models rather than to accurately predict PNC, more in-depth statistical analysis to test for representativeness was not performed.

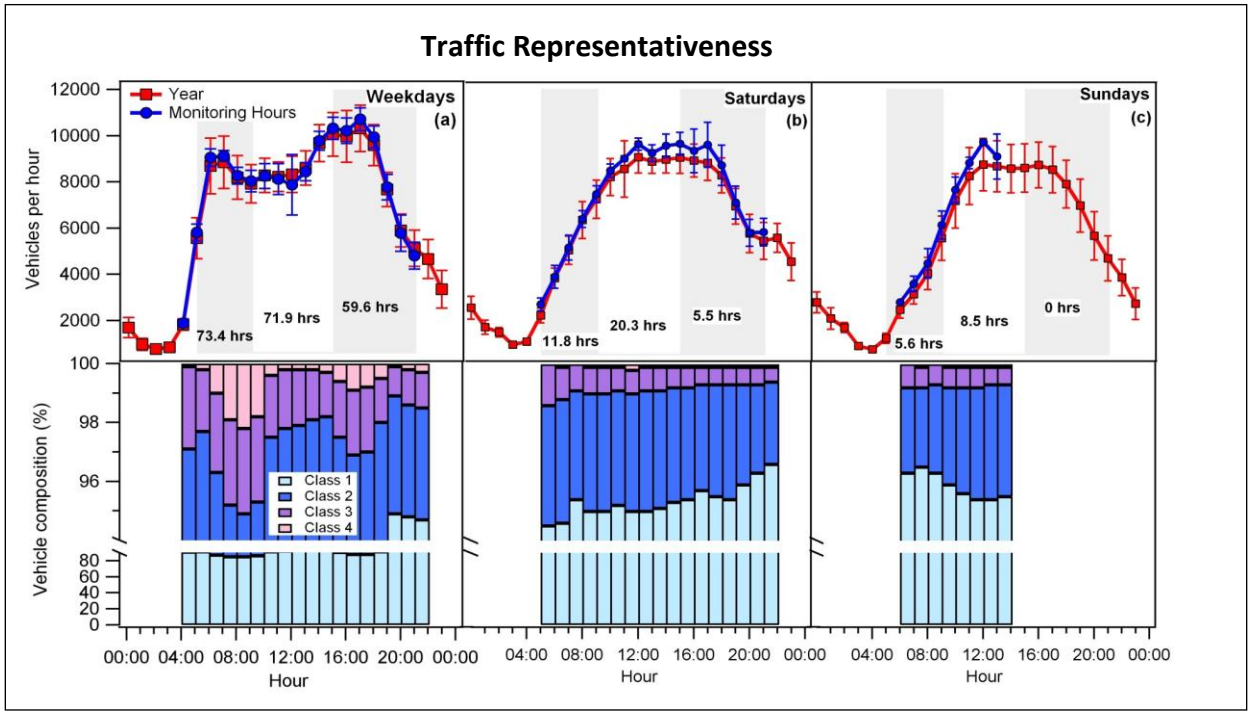

Figure 8: Comparison of traffic volume and composition during monitoring hours and over the entire study period. Class 1 is non-commercial gasoline-powered vehicles. Classes 2-4 are single-unit, single-trailer, and multi-trailer commercial vehicles, respectively. (Padró-Martínez et al., 2012)

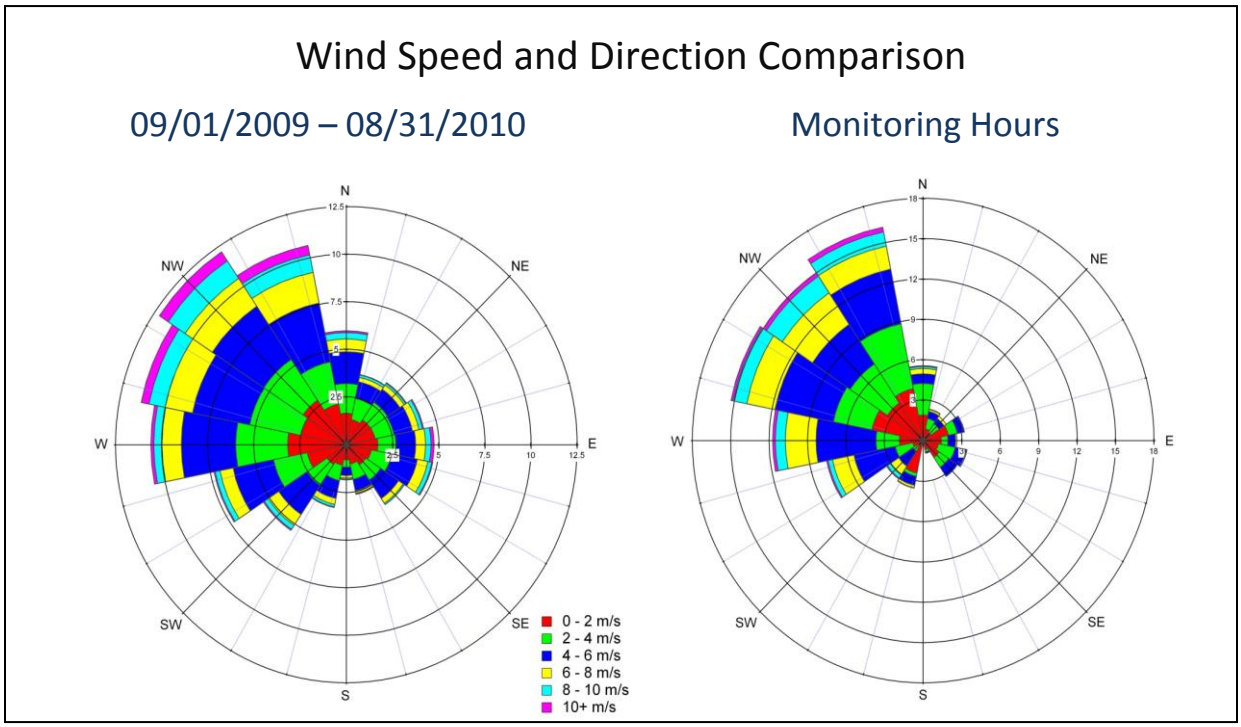

Figure 9: Wind speed was generally from the northwest during the entire year as well as during the monitoring period. East and southeast winds were underrepresented in the mobile monitoring dataset. (Padró-Martínez et al., 2012)

Using box plots, the data was grouped according to several different criteria such as wind speed, wind direction, season, and distance from the highway (Figures 10-11). From the box plots it was possible to identify several trends in the data. Variation with season and wind speed was as expected: winter concentrations were higher due to inefficient automobile combustion and residential heating, and wind speed varied inversely with pollutant concentration. From the wind-direction-segregated plot it was observed that concentrations were unusually high when the wind was out of the southeast, likely due to additional sources of pollution to the southeast of the study area (e.g. bus depot at Sullivan Square, highway intersection). From the box plots it is clear that many seasonal and wind-related variations are as great as the difference between near-highway and background concentrations (Figures 6-8). For example, median background concentrations (>1000 m from I-93) in the winter are as high as

near-highway (0-50 m) concentrations in the fall and summer From this we can infer that temporal factors such as wind and temperature will likely play a large role in a regression model for PNC.

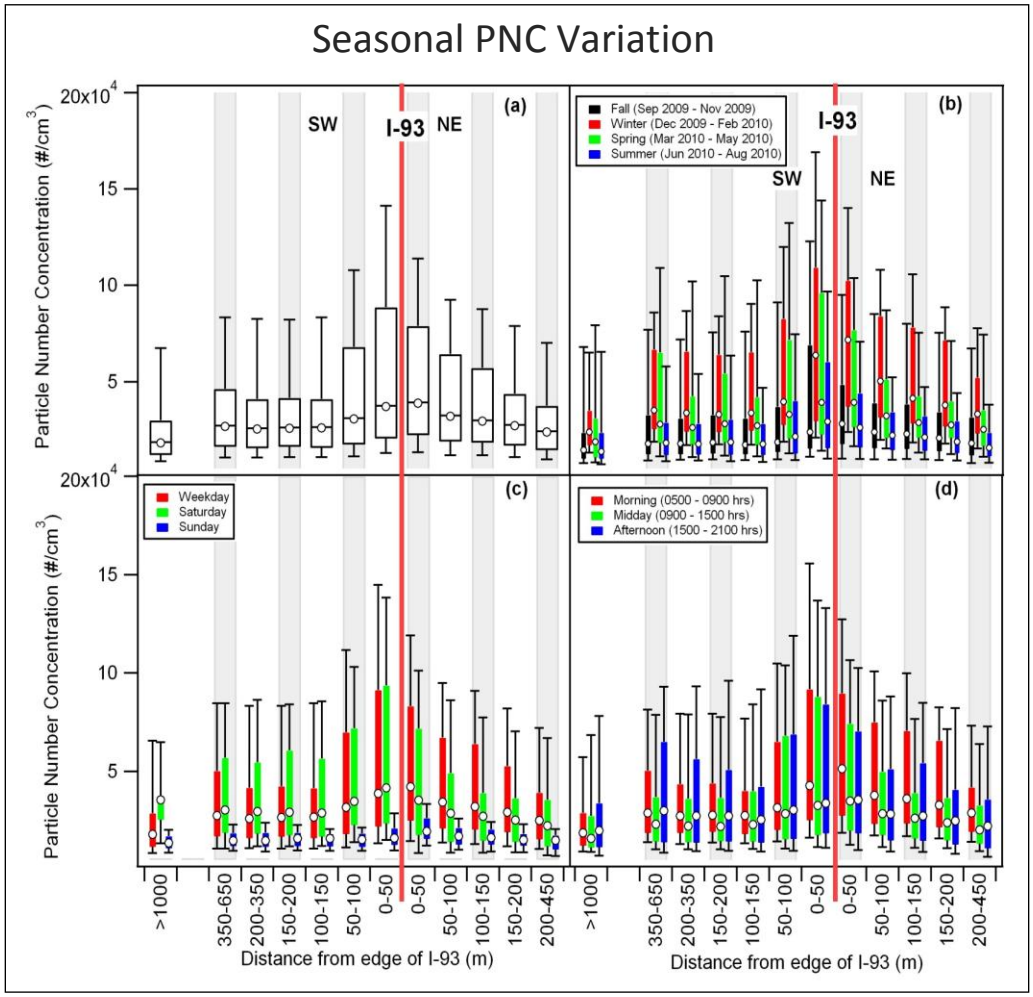

Figure 10: Plot of PNC averaged by distance to the highway (a) overall, (b) by season, (c) by weekday/weekend, and (d) by time of day. Note that often temporal variation is equal to or greater than the variation away from the highway. (Padró-Martínez et al., 2012)

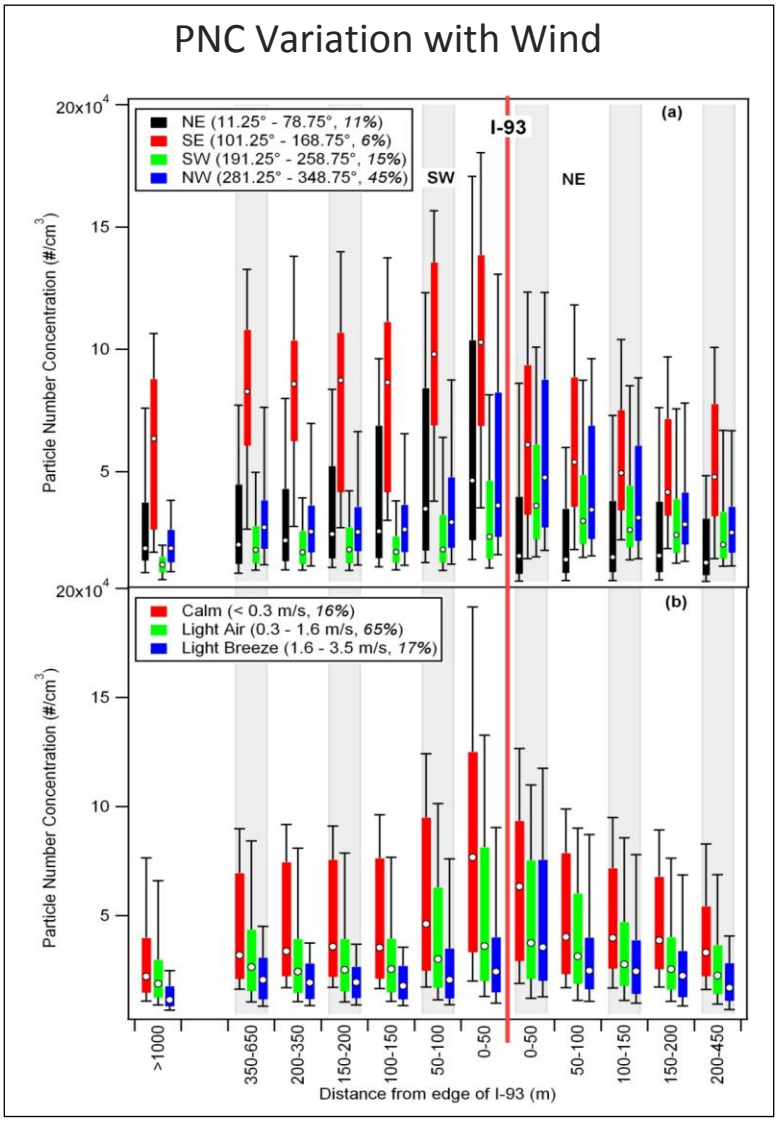

Figure 11: Plot of PNC averaged by distance to the highway and wind direction. (Padró-Martínez et al., 2012)

The autocorrelation plots for PNC measurements taken at the MAC (Figures 12-13) displayed a strong relationship between current and lagged measurements as well as strong periodicity, suggesting that the PNC measurements at any one location would be best modeled using an autoregressive model with cyclic components. This is expected, since the same atmospheric conditions generally prevail from one hour to another, thus producing similar rates of pollutant transport and dilution. Diurnal cycles of atmospheric stability and mixing height also have a strong effect on pollutant levels, which explains the cyclic nature of the ACF. Since the MAC data was collected at a stationary site (no spatial variation in the data) and is available as a continuous time series, it is possible to fully characterize its autocorrelation. While analyzing the MAC data does not give us any quantitative information about the mobile data, the fact that stationary measurements in the study area are temporally autocorrelated suggests that the mobile data will probably have the same high level of temporal autocorrelation. In addition, Figure 12 is very similar to the results found by Clifford et al. (2011), who also analyzed the ACF of PNC at a stationary site. This suggests that the ACF for PNC may be similar across different studies in some cases and further justifies the assumption that other parts of the study area would have the same ACF.

The regression of mobile data with 1-hour lagged concentrations had an adjusted R-squared of 0.22, which also supports the hypothesis that the dependent variable is autocorrelated. Based on these findings, the wind-variable regressions were run with and without the 1-hour-lagged concentration variable (lag_logPNC_Mean) to examine the effect of 1-hour-lag temporal autocorrelation on the model parameters, adjusted R^2, and coefficient of variation.

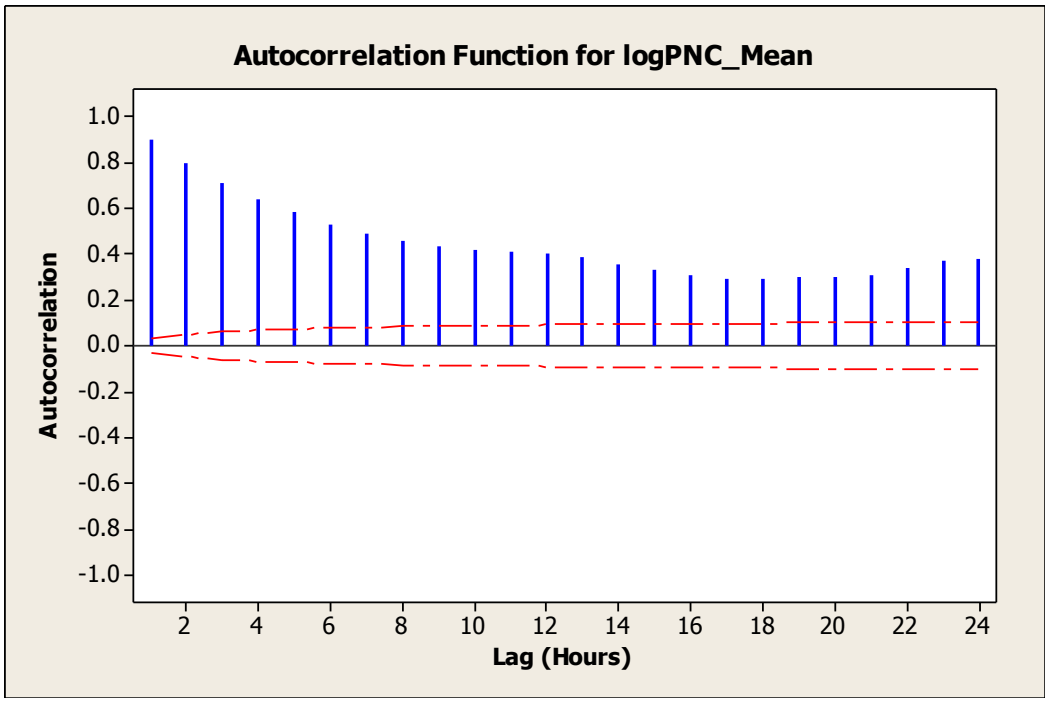

Figure 12: Temporal autocorrelation of MAC fixed-site PNC data with lags of 1-24 hours, 5% significance limits indicated by the dotted line. The MAC data is very highly autocorrelated, which suggests that mobile monitoring data is temporally autocorrelated as well.

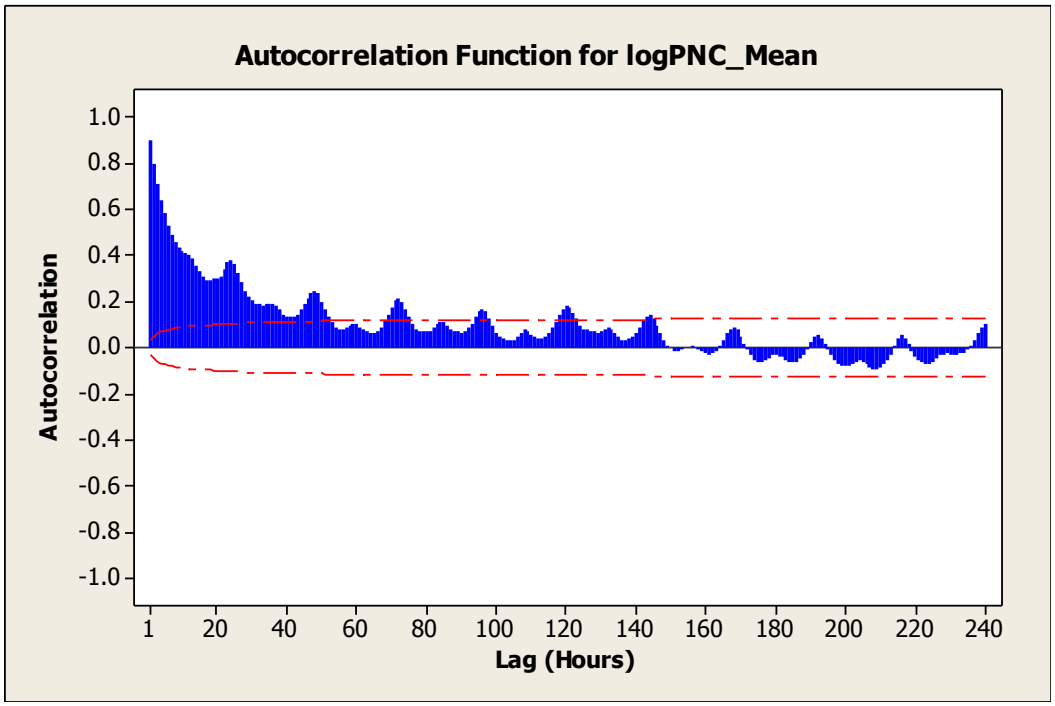

Figure 13: Temporal autocorrelation of the MAC fixed-site PNC data with lags up to 240 hours, 5% significance limits indicated by the dotted line. Strong autocorrelation is present for lags of less than 24 hours, and diurnal periodicity is also evident in the ACF.

3.2 Regression Results

Tables 2 and 3 contain parameters and summary statistics for two different base models (without and with 1-hour-lagged PNC) with a variety of wind variables. Because the models were not intended to predict PNC, they still have high coefficients of variation (models had a Cv of over 6, as compared to a Cv of 0.03 for ln(PNC) over one day of stationary TAPL monitoring).

Although none of the models had normally distributed residuals (probability plots were similar in shape to Figure 7), the non-normality was probably not enough to cause misrepresentation of the significance of model parameters, especially since t-ratios for the models were very large (over 200 in some cases). There was also a very slight downward trend in the residuals vs. predicted value (Figure 14), but the variance of the residuals remained constant. Again, this may cause some bias in the calculated model variances and significance tests but not enough to discredit them as a comparison tool, especially if there is a large difference in the two variances or t-ratios being compared. As expected, all VIFs were below 2, which indicated little to no correlation between model variables. Because several of the initial assumptions were violated, it will be impossible to trust parameter significance tests, model variance, or parameter estimates completely. However, most of the assumptions were violated only slightly, which means that the model results can still be meaningful. The models were compared using adjusted R^2 and all model parameters were examined to check whether they agreed with what would be expected from the theory.

In general, all model coefficients were significant ($p<0.0001$) unless otherwise noted. T-ratios were very large, probably due to the large size of the dataset. The t-tests for model parameters cannot be completely trusted due to the autocorrelation that is almost certainly present in the model error and the slight non-normality and heteroscedasticity of the residuals, but the fact that they were so high suggests that most parameters were indeed significant despite any bias that might be present.

Wind speed was found to be a very important factor in the variable formulation. Wind sectors and Fourier variables without wind speed included had a significantly lower adjusted R^2 than the same set of variables with wind speed included as an additional variable in the regression. Coefficients for the wind variable were almost always negative, reflecting the inverse relationship between PNC and wind.

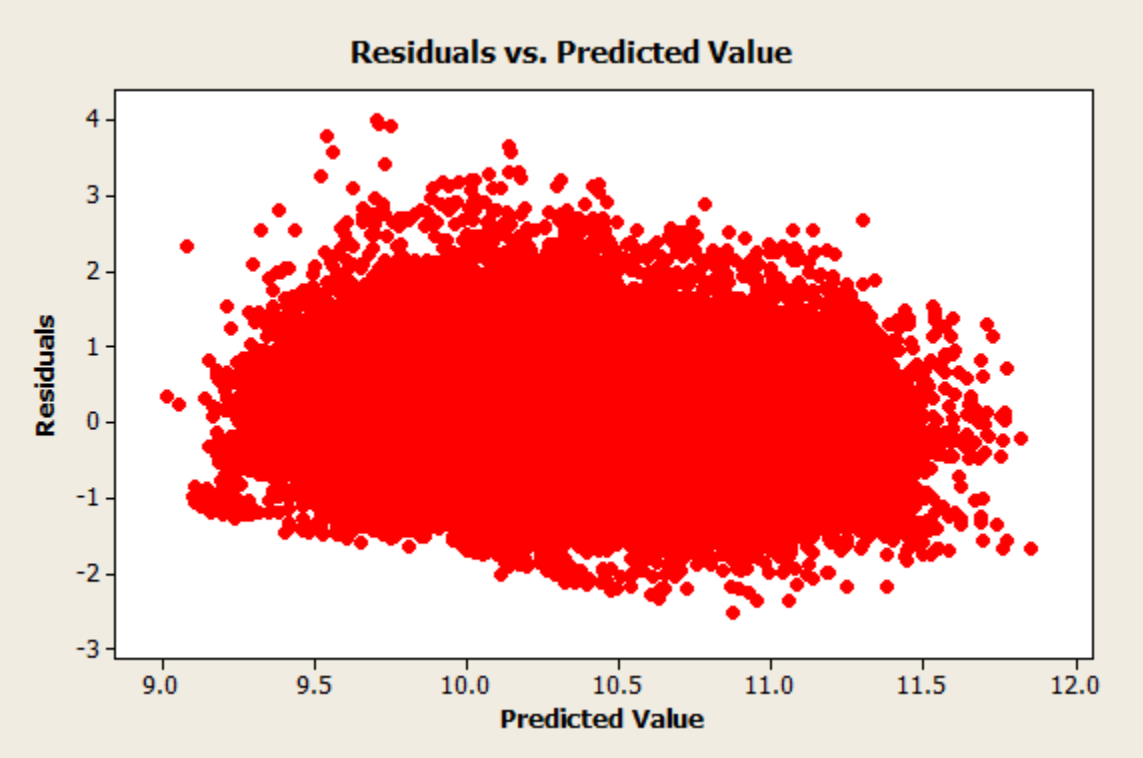

Figure 14: Scatterplot of residuals by predicted value for a regression with the 1-hour-lag ln(PNC) variable and highway-relative wind sectors. All residual plots were very similar between regressions that included the 1-hour lag, so this plot is representative of all regressions listed in Table 3.

3.2.1 Effect of the 1-hour-lagged PNC term

Although the 1-hour-lagged PNC term may help address autocorrelation issues in the model, it is not possible to assess temporal autocorrelation of the residuals directly. However, some basic analyses yielded results that suggested the lagged PNC term may have improved model performance.

The lagged PNC model (Table 3) only had 71086 observations, as opposed to the model without the lag (Table 2) which had 150914 observations. To investigate the effect that removing observations

had on the model results, a model with the same variable set (temperature, whether on a major road, wind sector, and wind speed) was run in both datasets. The full dataset had a slightly higher adjusted R^2 (0.36) and larger Cv (6.53) than the dataset including only times when lags could be computed (adjusted R^2=0.35, Cv=6.45). In addition, t values for every parameter were lower in the model from the smaller dataset. This may be due to the smaller dataset, but the variation in adjusted R^2 between the two models was so slight that it was determined they could be compared to each other without needing to account for the different sample size any further. Adding the lagged PNC term did not affect the normality of the residuals or the slight heteroscedasticity of the residuals with respect to the predicted value. Because PNC seems to be autocorrelated up to at least 24-hour lags, it would probably be necessary to add more lags greater than one hour to see a large effect on the model.

The lag term improved every model's adjusted R^2 and coefficient of variation, but did not change which models had the highest adjusted R^2 or affect the consistency of parameter estimates. Since the lag term may have reduced the autocorrelation in the residuals, the results with the lag term will be considered for the purposes of comparing wind variables.

Regression results for models run without 1-hour-lagged PNC term

Wind Variable	β_0	β_1	β_2	β_4	β_5	β_6	β_7	Adjusted R^2	Coefficient of Variation
Upwind, Downwind, ParallelSE, wsmps	11.28 (1865)	-0.018 (-175)	0.41 (117)	-0.065 (-14)	0.16 (33)	0.84 (111)	-0.092 (-118)	0.36	6.53
Sector1, Sector2, Sector3, wsmps	12.06 (1443)	-0.020 (-186)	0.41 (115)	-0.76 (-97)	-0.69 (-109)	-0.59 (-90)	-0.097 (-118)	0.35	6.57
Cosdir, Sindir, wsmps	11.65 (1747)	-0.022 (-203)	0.41 (115)	-0.30 (-82)	0.15 (53)	-0.081 (-97)		0.34	6.65
Upwind, Downwind, ParallelSE	10.92 (1998)	-0.019 (-175)	0.41 (111)	-0.016 (-3)	0.26 (53)	1.03 (132)		0.30	6.83
cosdir, sindir	11.47 (1743)	-0.023 (-210)	0.41 (111)	-0.38 (-100)	0.24 (89)			0.30	6.86
wsmps	11.40 (1831)	-0.018 (-179)	0.41 (110)	-0.11 (-140)				0.30	6.86
vnvws	10.94 (1924)	-0.018 (-175)	0.40 (105)	0.079 (88)				0.24	7.11
vd93dp	11.12 (1880)	-0.017 (-165)	0.31 (78)	-0.00052 (-83)				0.24	7.13
vwssq	11.14 (1834)	-0.018 (-178)	0.41 (106)	-0.0067 (-73)				0.23	7.17
vdp	10.74 (1622)	-0.017 (-162)	0.40 (104)	0.23 (73)				0.23	7.17
vws	11.10 (1747)	-0.018 (-173)	0.41 (104)	-0.028 (-44)				0.21	7.25
vd93ws	11.02 (1879)	-0.017 (-162)	0.37 (92)	-0.00027 (-38)				0.21	7.26
vwsd93dp	10.99 (1889)	-0.018 (-168)	0.40 (102)	0.0012 (4)				0.20	7.29

Table 2: Parameters for regressions following the model form given in Section 2.5 with t-ratios shown underneath in parentheses. β_0 is the model intercept, and β_1, β_2, and β_4—β_7 are the intercepts for TempF, onmaj, and the wind variables in the order they are listed in the left-hand column. These regressions did not contain a 1-hour-lagged PNC term, so β_3=0 for every model in the table. For definitions of variables, see Table 1 or Appendix 2.

Regression results for models run with 1-hour-lagged PNC term

Wind Variable	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	Adjusted R^2	Coefficient of Variation
Upwind, Downwind, ParallelSE, wsmps	8.39	-0.014	0.28	0.27	-0.080	0.15	0.60	-0.091	0.41	6.15
	(231)	(-90)	(57)	(85)	(-13)	(23)	(55)	(-86)		
Sector1, Sector2, Sector3, wsmps	8.87	-0.015	0.28	0.28	-0.60	-0.46	-0.35	-0.098	0.40	6.17
	(232)	(-97)	(56)	(86)	(-56)	(-53)	(-39)	(-88)		
Cosdir, Sindir, wsmps	8.68	-0.017	0.28	0.27	-0.25	0.069	-0.086		0.39	6.22
	(229)	(-107)	(57)	(84)	(-47)	(18)	(-76)			
Upwind, Downwind, ParallelSE	7.85	-0.014	0.27	0.28	-0.032	0.22	0.78		0.35	6.46
	(209)	(-85)	(52)	(85)	(-5)	(32)	(69)			
Cosdir, Sindir	8.38	-0.018	0.28	0.28	-0.33	0.16			0.34	6.47
	(214)	(-106)	(53)	(82)	(-59)	(43)				
wsmps	8.17	-0.013	0.27	0.30	-0.10				0.37	6.34
	(220)	(-94)	(52)	(92)	(-96)					
vnvws	7.56	-0.013	0.25	0.31	0.065				0.31	6.62
	(198)	(-87)	(48)	(92)	(51)					
vd93dp	7.66	-0.012	0.22	0.31	-0.00037				0.30	6.69
	(197)	(-79)	(41)	(90)	(-32)					
vwssq	7.70	-0.013	0.25	0.32	-0.0067				0.32	6.58
	(202)	(-90)	(48)	(94)	(-58)					
vdp	7.41	-0.012	0.25	0.31	0.17				0.30	6.66
	(193)	(-79)	(47)	(91)	(40)					
vws	7.59	-0.013	0.25	0.33	-0.031				0.30	6.67
	(197)	(-86)	(46)	(95)	(-38)					
vd93ws	7.47	-0.012	0.24	0.32	-0.000018**				0.29	6.74
	(192)	(-80)	(45)	(93)	(-1.34)					
vwsd93dp	7.47	-0.012	0.24	0.32	0.0010**				0.29	6.74
	(193)	(-81)	(45)	(94)	(2.78)					

Table 3: Parameters for regressions following the model form given in Section 2.5 with t-ratios shown underneath in parentheses. β_0 is the model intercept. β_1, β_2, β_3, and β_4—β_7 are the coefficients for TempF, onmaj, lag_logPNC_Mean, and the wind variables in the order they are listed in the left-hand column. The variable lag_logPNC_Mean is a 1-hour-lagged PNC term, which caused a slight increase in the adjusted R^2, as well as a decrease in the coefficient of variation. For definitions of variables, see Table 1 or Appendix 2. (** model coefficient not significant; p>0.001)

3.2.2 Performance of vector variables

In general, the vector-based wind variables performed poorly compared to other wind variables (Table 3). Since they all produced a lower adjusted R^2 than wind speed alone, it is very likely that there is some consistent reason why using a vector-based variable for this particular dataset and model is not as effective. One possible reason for the poor performance of the vector variables is the fact that they are very descriptive in their interpretation of wind-based particle motion, while the actual wind field and its effect on particles is too chaotic to be characterized by a theoretical relationship with a single wind speed and direction. More averaged and generalized wind variables may have performed better because the wind at the site is only characterized generally. To further characterize the complex urban wind field and possibly improve the performance of the vector variables, multiple stations could be placed within the monitoring area and the wind direction and speed could either be interpolated over the entire area or each data point could be assigned a wind speed/direction from the nearest station.

3.2.3 Performance of wind sectors

The highway-relative wind sectors produced the highest adjusted R^2 and the lowest coefficient of variation out of all the wind variables in each model (Table 3), and the conventional wind sectors produced similar results, with a slightly lower adjusted R^2 and higher coefficient of variation. For the highway-relative wind sectors, the highest coefficient was for the variable corresponding to wind parallel to the highway from the southeast. The coefficients for conventional Sectors 1, 2, and 3 were negative, which indicated that Sector 4 (wind from the southeast) produced the highest PNC conditions. This is may be due to the many possible sources of PNC located east and southeast of the study area.

3.2.4 Performance of first-order Fourier approximation

The Fourier variables performed similarly to the wind sector variables, with a slightly lower adjusted R^2 (Table 3). The coefficients from the regression including wind speed correspond to a sinusoidal curve having a maximum at 120°, or southeasterly wind direction. This agrees with intuition, since most major

sources of pollution are in the southeastern part of the study area, including I-93 as it turns to the south leaving Somerville and the intersection with Mass. Route 28 (Figure 2). The Fourier variables might be further improved by adding in higher-order terms to better approximate the complex response of PNC to wind direction.

4.0 CONCLUSIONS

The results from this study will help guide the choice of a wind variable in the regression model of PNC that will aid in assessing exposure to UFP for the study participants of CAFEH. Although the results of this paper are specific to a particular dataset, many of the methods discussed herein can be used to determine the best wind variable for other regression models of PNC. It is important to realize that the way wind is represented in a linear regression model is very important, and that statistical insignificance of the wind variable may be a product of incorrect variable form, rather than a true lack of relationship between PNC and wind.

4.1 Optimal Wind Variable

For this particular dataset and regression method, it was determined that highway-relative wind sectors are the best choice of variable to represent the wind effect on PNC, since they yielded the highest adjusted R^2 and their parameters agreed with expectation. Other variable types, such as conventional wind sectors or Fourier-based variables, also have great potential for incorporating wind information into a regression model of PNC. Further research should investigate the benefits of using smaller wind sectors and Fourier-based variables of higher order, which will allow for a more fine-grained characterization of the effects of nearby sources of pollution and may increase the explanatory power of the wind variable even further. However, when a major source of a pollutant is within the study area, rather than outside of it, better performance may be seen for wind variables which account for wind direction relative to that source's position, as the highway-relative wind sectors did in this study.

Care must be taken not to make regression variables more finely-resolved than the data they are based on. The vector variables in this study performed very poorly, most likely due to the fact that (1) there was a large amount of error inherent in the wind estimate since only one weather station was

used and (2) the variables were so precisely defined that the error in the wind data was able to propagate into the final model, causing a sensitivity to error similar to that observed in dispersion models. If the wind field is better characterized in future studies by using more weather stations, these variables may perform better.

4.2 Autocorrelation of the Data

A concerted effort must be made to address autocorrelation in the model to comply with the assumptions of linear regression. Because it is impossible to directly address temporal autocorrelation in the model due to the limitations of the dataset, more indirect methods should be used, such as including variables that reflect the periodicity or modeling the difference between mobile monitoring measurements and a central monitoring cite. The ACF of PNC in this study was similar to that observed in another study, which may indicate that PNC is always highly temporally autocorrelated. Any model of PNC with a fine temporal resolution (daily or hourly) should test for autocorrelation of the data and take measures accordingly.

For mobile monitoring studies, a priority should be designing methods that allow the temporal autocorrelation of the data to be calculated to more lags. This can be achieved by increasing the regularity of the monitoring schedule so that the same locations are measured every hour. Additionally, it is very important to have a reliable fixed site for a mobile monitoring study, since it serves as a measure of temporal autocorrelation at higher lags and including it in the model may sometimes be the only way to eliminate autocorrelation of the residuals.

5.0 ACKNOWLEDGMENTS

The author would like to thank her thesis committee professors John Durant and Richard Vogel, as well as her graduate student advisor Allison Patton for their invaluable help and advice. The author would also like to gratefully acknowledge the CAFEH mobile monitoring team: Jeffrey Trull, Eric Wilburn, Piers MacNaughton, Kevin Stone, and Jessica Perkins. Thanks also to Doug Brugge, Wig Zamore, and all the CAFEH staff and investigators. This research was supported by Tufts Summer Scholars, the Tufts Undergraduate Research Fund, and the Cataldo Scholars program. CAFEH is supported by a grant from the National Institute of Environmental Health Sciences (Grant No. ES015462), the Tufts University Tisch College, and the Tufts Community Research Center.

REFERENCES

Arain, M. a., Blair, R., Finkelstein, N., Brook, J. R., Sahsuvaroglu, T., Beckerman, B., Zhang, L., et al. (2007). The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. *Atmospheric Environment*, *41*(16), 3453-3464.

Briggs, D. (2005). The role of GIS: coping with space (and time) in air pollution exposure assessment. *Journal of Toxicology and Environmental Health. Part A*, *68*(13-14), 1243-61.

Brugge, D., Durant, J. L., & Rioux, C. (2007). Near-highway pollutants in motor vehicle exhaust: a review of epidemiologic evidence of cardiac and pulmonary health risks. *Environmental Health*, *6*, 23.

Brunekreef, B., Janssen, N. A. H., de Hartog, J., Harssema, H., Knape, M., & van Vliet, P. (1997). Air pollution from truck traffic and lung function in children living near motorways. *Epidemiology*, *8*(3), 298–303.

Clifford, S., Low Choy, S., Hussein, T., Mengersen, K., & Morawska, L. (2011). Using the Generalised Additive Model to model the particle number count of ultrafine particles, *Atmospheric Environment*, *45*(32), 5934-5945.

Durant, J. L., Ash, C. A., Wood, E. C., Herndon, S. C., Jayne, J. T., Knighton, W. B., Canagaratna, M. R., et al. (2010). Short-term variation in near-highway air pollutant gradients on a winter morning. *Atmospheric Chemistry and Physics*, *10*(17), 8341-8352.

Fruin, S., Westerdahl, D., Sax, T., Sioutas, C., & Fine, P.M. (2008). Measurements and predictors of on-road ultrafine particle concentrations and associated pollutants in Los Angeles. *Atmospheric Environment 42*, 207-219.

Gauderman, W. J., Vora, H., McConnell, R., Berhane, K., Gilliland, F., Thomas, D., Lurmann, F., et al. (2007). Effect of exposure to traffic on lung development from 10 to 18 years of age: a cohort study. *The Lancet*, *369*(9561), 571-577.

Hagler, G., Thoma, E., & Baldauf, R. (2010). High-resolution mobile monitoring of carbon monoxide and ultrafine particle concentrations in a near-road environment. *Journal of the Air & Waste Management Association*, *60*(3), 328-336.

Helsel, D. R. & R. M. Hirsch (2002). Statistical Methods in Water Resources. *Techniques of Water-Resources Investigations* (Book 4, Chapter A3). U.S. Geological Survey.

Hinds, W. C. (1999). *Aerosol Technology* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

Hoek, G., Beelen, R., Hoogh, K. de, & Vienneau, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, *42*, 7561-7578.

Hwang, B.-F., Lee, Y.-L., Lin, Y.-C., Jaakkola, J. J. K., & Guo, Y. L. (2005). Traffic related air pollution as a determinant of asthma among Taiwanese school children. *Thorax*, *60*(6), 467-473.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., et al. (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology*, *15*(2), 185-204.

Klems, J. P., Pennington, M. R., Zordan, C. A. & Johnston, M. V. (2010). Ultrafine particles near a roadway intersection: Origin and apportionment of fast changes in concentration. *Environmental Science & Technology, 44*(20), 7903-7907.

McConnell, R., Berhane, K., Yao, L., Jerrett, M., Lurmann, F., Gilliland, F., Kunzli, N., et al. (2006). Traffic, susceptibility, and childhood asthma. *Environmental Health Perspectives*, *114*(5), 766-772.

Mølgaard, B., Hussein, T., Corander, J., & Hämeri, K. (2011). Forecasting size-fractionated particle number concentrations in the urban atmosphere. *Atmospheric Environment*, 1-9. Elsevier Ltd.

Oberdörster, G., Oberdörster, E., & Oberdörster, J. (2005). Nanotoxicology: An emerging discipline evolving from studies of ultrafine particles. *Environmental Health Perspectives*, 113(7), 823-839.

Padró-Martínez, L., Patton, A., Trull, J.B., Zamore, W., Brugge, D., Durant, J.L., (2012). Submitted to *Atmospheric Environment*. Mobile monitoring of spatial and temporal variation of traffic-related air pollution in a near-highway urban neighborhood over the course of a year.

Seaman, N. (2000). Meteorological modeling for air-quality assessments. Atmospheric environment, 34, 2231-2259.

Sioutas, C., Delfino, R. J., & Singh, M. (2005). Exposure assessment for atmospheric ultrafine particles (UFPs) and implications in epidemiologic research. *Environmental Health Perspectives*, 113(8), 947-955.

Su, J. G., Brauer, M., Ainslie, B., Steyn, D., Larson, T., & Buzzelli, M. (2008). An innovative land use regression model incorporating meteorology for exposure analysis. *The Science of the Total Environment*, *390*(2-3), 520-529.

van Vliet, P., Knape, M., de Hartog, J., Janssen, N., Harssema, H., & Brunekreef, B. (1997). Motor vehicle exhaust and chronic respiratory symptoms in children living near freeways. *Environmental Research*, *74*(2), 122-132.

Zhu, Y., Hinds, W. C., Kim, S., & Sioutas, C. (2002). Concentration and size distribution of ultrafine particles near a major highway. *Journal of the Air & Waste Management Association (2002)*, *52*(9), 1032-1042.

Zwack, L. M., Paciorek, C. J., Spengler, J. D., & Levy, J. I. (2011). Modeling spatial patterns of traffic-related air pollutants in complex urban terrain. *Environmental Health Perspectives*, 119(6), 852-859.

APPENDICES

APPENDIX 1: SAS CODE TO CALCULATE 1-HOUR LAGS OF MOBILE MONITORING DATA

```
libname Caitlin 'insert filepath here';

****INITIALIZE DATASET. DEFINE ABSTIME VARIABLE****
data thisreg;
        set Caitlin.regready;
        *abstime is the number of hours since the first hour of 1/1/2009;
        abstime=(year-2009)*365*24+(day-1)*24+hour;
run;

****SORT DATA BY LOCATION, THEN BY TIMESTAMP****
proc sort data=Caitlin.regready;
        by sgmt_ID abstime;
run;

****CREATE A NEW DATASET WHICH CONTAINS 1-HOUR LAGS****
data Caitlin.regready_new;
        set Caitlin.regready (where=(logpnc_mean ne .));
        *calculate 1-hour lagged concentration variable if data is available;
        If (abstime-lag(abstime)=1 and sgmt_ID=lag(sgmt_ID)) then lag_logpnc_Mean=lag(logpnc_Mean);
        if ((abstime-lag(abstime) ne 1) or (sgmt_ID ne lag(sgmt_ID))) then lag_logpnc_Mean=.;
run;
```

APPENDIX 2: SAS CODE WIND VARIABLE DEFINITIONS

```
libname Caitlin 'insert filepath here';

data Caitlin.regready_new;
        set Caitlin.regready;


        ****WIND SECTOR DEFINITIONS****
        if (0<wdir<=90) then Sector1=1;
        else if wdir=. then Sector1=.;
        else Sector1=0;
        label Sector1="wind from the northeast";

        if (270<wdir<=360) then Sector2=1;
        else if wdir=. then Sector2=.;
        else Sector2=0;
        label Sector2="wind from the northwest";

        if (180<wdir<=270) then Sector3=1;
        else if wdir=. then Sector3=.;
        else Sector3=0;
        label Sector3="wind from the southwest";

        if (90<wdir<=180) then Sector4=1;
        else if wdir=. then Sector4=.;
        else Sector4=0;
        label Sector4="wind from the southeast";


        ****HIGHWAY-RELATIVE WIND SECTOR DEFINITIONS****
        if (Hwy_Side="West" and (157.5<wdir<292.5)) then Upwind=1;
        else if (Hwy_Side="East" and (0<wdir<112.5 or 337.5<wdir<360)) then Upwind=1;
        else if wdir=. then Upwind=.;
        else Upwind = 0;
        label Upwind = "upwind of highway";

        if (Hwy_Side="West" and (0<wdir<112.5 or 337.5<wdir<360)) then Downwind=1;
        else if (Hwy_Side="East" and (157.5<wdir<292.5)) then Downwind=1;
        else if wdir=. then Downwind=.;
        else Downwind=0;
        label Downwind = "downwind of highway";

        if (112.5<wdir<157.5 or 292.5<wdir<337.5) then Parallel = 1;
        else if wdir=. then Parallel=.;
        else Parallel =0;
        label Parallel = "wind parallel to highway";
```

```
        if (Parallel=1 and wdir<180) then ParallelSE=1;
        else if Parallel=. then ParallelSE=.;
        else ParallelSE=0;
        if (Parallel=1 and wdir>180) then ParallelNW=1;
        else if Parallel=. then ParallelNW=.;
        else ParallelNW=0;


        ****VECTOR VARIABLE DEFINITIONS****
        vdp=cos(2*constant('pi')*(wdir-Hwy_Direction)/360)+1.0;
        vd93ws=disti93/(wsmps10m*vdp+1.1);
        vwsd93dp=wsmps10m*vdp/(disti93+0.1);
        vd93dp=disti93/(vdp+1.1);
        vws=wsmps10m*vdp;
        vwssq=wsmps10m*wsmps10m*vdp;
        vnvws=(1/(wsmps10m+0.1))*vdp;


        ****FOURIER VARIABLE DEFINITIONS****
        cosdir=cos(2*constant('pi')*wdir/360);
        sindir=sin(2*constant('pi')*wdir/360);

;
run;
```