COHERENT APPROACH FOR MODELING AND NOWCASTING HOURLY NEAR-ROAD BLACK CARBON CONCENTRATIONS IN SEATTLE, WASHINGTON

Runze Yu, Ph.D.
Urban Transport Consultant
China and Mongolia Sustainable Development Unit
East Asia and Pacific Region, World Bank
17th Floor, China World Office 2
No. 1 Jianguomenwai Avenue
Beijing, P.R. China 100004
Tel: 18610906057, Email: runze@uw.edu

Xiaoyue Cathy Liu, E.I.T.
Ph.D. Candidate, Graduate Research Assistant
Department of Civil and Environmental Engineering, University of Washington
Box 352700, Seattle, WA 98195-2700
Tel: (281) 760-9768, Email: liuxy@uw.edu

and

*Yinhai Wang, Ph.D.
Professor
Department of Civil and Environmental Engineering, University of Washington
Box 352700, Seattle, WA 98195-2700
Tel: (206) 616-2696, Fax: (206) 543-1543, Email: yinhai@uw.edu

Submitted for Presentation at the 92nd Transportation Research Board Annual Meeting
Washington, D.C., January, 2013

Date Submitted: November 12th, 2012

Word Count: 4,369 words + (4 figures * 250 words) + (6 tables*250 words) = 6,869 words

*Corresponding author
ABSTRACT

With growing awareness for near-road air pollution and an increasing population of pedestrians, it is imperative to “nowcast” near-road air quality conditions to the general public, which necessitates the efforts of building hourly predictive models with ease of use and satisfactory accuracy. This study demonstrates a coherent approach to model the hourly near-road Black Carbon (BC) concentrations with on-road factors and meteorological conditions using datasets from two urban sites in Seattle, Washington. With Bayesian Model Averaging (BMA) method, the optimal set of regressors is determined. Three different model structures are further developed and compared by goodness-of-fit. An innovative approach is proposed to translate wind direction from numerical values to categorical variables with statistical significance. By modeling the autocorrelation within the BC time series using AR(1) component, the developed approach yields a satisfying prediction accuracy. The conditional heteroscedasticity and heavy-tailed distribution within the residuals are successfully identified and modeled by the General Auto Regressive Conditional Heteroscedasticity (GARCH) model, which provides valuable insights to the interpretation of prediction results. The approach demonstrated in this study presents an efficient and valuable solution to selecting and fine-tuning a nowcasting model to be implemented onto online platforms for near-road BC nowcasting. A comparison between the two study sites also reveals the effectiveness of freight regulation for mitigating environmental impacts from the heavy truck fleet.

KEYWORDS: Black Carbon, Freight, Time Series Analysis, Nowcast
INTRODUCTION

Numerous studies have identified high air pollutant concentrations in the ambient environment near major roadways, to which the exhaust emitted by the on-road traffic accounts for a major contributor (1). Traffic related air pollution is found to have direct threats to public health (2, 3 and 4). With a population of over 40 million living within 300 feet (91m) of major transportation infrastructures (5) and approximately 250 million vehicles on the roads, air pollution from road traffic has imposed great challenges to the sustainability of the modern society.

The U.S. Environmental Protection Agency (EPA) therefore has strengthened the monitoring practice and national standard for near-road air pollutants. Investigating near-road human exposure requires a finer temporal resolution for air quality monitoring and forecasting, as the traditional daily average reporting fails to capture the vast temporal and spatial span of human activities throughout the day. Also the concentrations of many air pollutants vary to a great extent, e.g. peak hours’ concentrations far exceed the daily average (6). Short-term exposure has been found to have more profound public health impact (7) than the daily average exposure, which highlights the practical importance of hourly forecasts. Nowcasting air pollutant concentrations often focuses on specific areas and provides short-period forecasts (8). It has been widely executed in weather forecasting and economics prediction. The time window ranges from the next one to 12 hours depending on different scopes of needs. In the context of this study, nowcasting would allow the general public to assess immediate air quality conditions and to plan their activities along the roadways (e.g. jogging or biking) accordingly.

The nowcasting result of air quality can be delivered to the public via Geographic Information System (GIS) based online platforms. Previous research (9, 10, 11 and 12) has demonstrated that those systems have the capability of not only housing a diverse amount of datasets, but also broadcasting forecasts from backstage computational engine in real-time. This requires the system to optimally select impact factors, adopt computationally inexpensive model structures, and quickly produce satisfying predictions.

A variety of statistical models applied by previous studies (6 and 13) have brought prospects towards building such a computational engine. Sfetsos and Vlachogiannis (14) applied Local Models with Clustering Algorithms (LMCA) and Hybrid Clustering Algorithm (HCA) to predicting hourly PM$_{10}$ concentration. It was found that the linear approach marginally outperformed Artificial Neural Network (ANN) approaches. Ballester, et al. (15) built pure predictive models for one-day-ahead forecasts of the hourly ozone concentration across three sites in Spain. Methods applied include autoregressive-moving average with exogenous inputs (ARMAX), Multilayer Perceptron (MLP) and FIR neural networks. For ARMAX models, R-squared values ranged from 0.75 to 0.8 and IAs ranged from 0.9 to 0.93, which were on par with both neural networks methods. Chen, et al. (16) modeled the univariate hourly ozone time series, using nonlinear phase space model. The model was then compared with optimum autoregressive model (a linear one) and R-squared values were 0.80 and 0.70 respectively. For one-hour-ahead prediction, the phase space model achieved prediction accuracy of 0.77, and 0.74 for the autoregressive model. Hrust, et al. (17) compared MLP neural networks models and linear models for forecasting hourly concentrations of NO$_2$, O$_3$, CO and PM$_{10}$ in city of Zagreb, Croatia, where the developed model triumphed over the linear model. While exploring different modeling methods, most of previous studies did not sufficiently emphasize the modeling process, including data preparation, variable selection, and model structure, which would have considerable influence on accurately capturing the underlying nature of the governing relationship between air...
pollutant concentrations and impacting factors; and if not conducted carefully, would weaken the model performance.

Besides the modeling process, another issue emerges on the data side. Modeling the relationship between hourly concentrations and on-road factors requires traffic data with high resolution and enriched information. Traditional traffic counters are able to capture general traffic fluctuations (18 and 19) yet fail to provide vehicle classification information, which is desirable for discovering the relationship between air pollutants and diesel fuel vehicles. Some researchers (20 and 21) utilized video data to resolve this issue as detailed traffic information can be retrieved from videos (22). Other researchers (23) relied on the Weigh-In-Motion (WIM) systems for high-resolution truck data. However, video data requires large storage space and computing power for processing, and thus can only be collected for a short period of time for validation purposes. Also, WIM stations are generally not located in urban areas which cannot meet the needs for urban air pollution modeling. Additionally, previous studies have generally identified that meteorological conditions have a direct impact on observed concentrations (6 and 24). Of various meteorological parameters, wind direction significantly affects hourly concentration modeling and forecasting (25).

In view of modeling of hourly air pollutant concentrations, this paper presents a coherent methodological approach to performing BC concentrations nowcasting on the basis of collected datasets at two urban sites in Seattle, Washington. BC refers to the soot particles produced during incomplete combustion. As a primary component of fine particular matter (PM$_{2.5}$), it has been found associated with public health risks and global warming effects (26). The objectives of this study are: 1) to address the challenging issues identified from previous research regarding air pollutant concentration modeling process, including variable selection, model structure, and residual analysis; 2) to quantitatively examine the impact of on-road traffic and meteorological conditions on hourly near-road BC concentrations, particularly by proposing an innovative method to incorporate wind fields into the model; and 3) to develop a computationally efficient model with desirable prediction accuracy for hourly BC concentration nowcasting implementable on web-based platforms.

**DATA AND METHODOLOGIES**

**Datasets**

Two near-road locations are chosen as the study sites in the urban area of Seattle as shown in Figure 1. The Industrial/Arterial site (Duwamish site) is located in the heart of the Duwamish valley. State Route 99 (SR-99) was the major roadway, also serving as the freight corridor connecting port terminals, rail yards, local manufacturers, and distribution warehouses in the area. The daily traffic at this site is approximately 50,000 vehicles with a significant number of diesel trucks observed. This site is considered as a location with the densest fine particle concentration in the Washington State (27). The Freeway site (Olive site) is in downtown Seattle (Olive Way and Boren Ave) in close proximity to Interstate 5 (I-5). Around 200,000 vehicles pass this site every day, with the majority of traffic as passenger cars.
Hourly BC concentrations and meteorological conditions were recorded by two nearby air quality monitoring stations on site. For the Duwamish site, the station is 100 feet (30 m) away from the SR-99. The Olive site’s station is located 65 feet (20 m) away from the I-5. Both stations belong to the air quality monitoring network in the Puget Sound region, managed and operated by Puget Sound Clean Air Agency (PSCAA) and Washington State Department of Ecology (WSDOE). Besides the BC concentrations captured by Aethalometer™, hourly meteorological data including ambient temperature, wind direction, and wind speed were also collected from a 10 meter tower. The PSCAA reviews, validates, and archives the data. Traffic data at the Duwamish site were collected by 12 Sensys™ VSN24-F detectors installed on SR-99. The Olive site’s traffic data were gathered by loop detectors and acquired from the Washington State Department of Transportation (WSDOT). For both sites a dual-detector configuration was activated to collect vehicle length information to infer vehicle classifications of passenger cars and heavy trucks.

Table 1 shows descriptive statistics of the data being collected. Six hours’ BC concentrations are missing out of a total of 2,256 hours at the Duwamish site. After aligning different datasets by timestamp, imputation is processed using linear interpolation between closest available data points. For the Duwamish site, 2,256 observations are used as training dataset and 672 used as testing dataset (1620/480 for the Olive site). For each variable, the upper and lower rows give the statistics for the training and testing datasets, respectively. Vehicle speed data for the Olive site is dropped from analysis due to data quality issue.
### TABLE 1 Statistics of Measured Values

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Unit</th>
<th>Sample Size</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Duwamish Site</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training dataset</td>
<td>from 00:00 March 9th to 23:00 June 10th, 2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing dataset</td>
<td>from 00:00 September 1st to 23:00 September 28th, 2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black Carbon</td>
<td>Micrograms per cubic meter</td>
<td>2256/672</td>
<td>1.1</td>
<td>0.8</td>
<td>0</td>
<td>11</td>
<td>1.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.7</td>
<td>1.2</td>
<td>0</td>
<td>14</td>
<td>1.76</td>
</tr>
<tr>
<td>Wind Direction</td>
<td>Degrees from the north</td>
<td></td>
<td>213.6</td>
<td>201.0</td>
<td>1</td>
<td>360</td>
<td>89.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>235.4</td>
<td>229.0</td>
<td>2</td>
<td>359</td>
<td>87.81</td>
</tr>
<tr>
<td>Wind Speed</td>
<td>Miles per hour</td>
<td></td>
<td>4.8</td>
<td>4.3</td>
<td>1</td>
<td>18</td>
<td>2.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>3.7</td>
<td>1</td>
<td>13</td>
<td>2.21</td>
</tr>
<tr>
<td>Temperature</td>
<td>Degrees Fahrenheit</td>
<td>2256/672</td>
<td>52.4</td>
<td>51.0</td>
<td>30</td>
<td>87</td>
<td>10.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>62.6</td>
<td>62.0</td>
<td>47</td>
<td>85</td>
<td>6.67</td>
</tr>
<tr>
<td>Speed</td>
<td>Miles per hour</td>
<td></td>
<td>40.7</td>
<td>41.5</td>
<td>28</td>
<td>51</td>
<td>4.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>41.3</td>
<td>43.0</td>
<td>27</td>
<td>49</td>
<td>4.77</td>
</tr>
<tr>
<td>CAR Volume</td>
<td>Vehicles per hour</td>
<td></td>
<td>2038.0</td>
<td>1746.5</td>
<td>129</td>
<td>5798</td>
<td>1445.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2093.0</td>
<td>1844.0</td>
<td>182</td>
<td>5735</td>
<td>1453.00</td>
</tr>
<tr>
<td>TRUCK Volume</td>
<td>Vehicles per hour</td>
<td></td>
<td>58.0</td>
<td>15.0</td>
<td>0</td>
<td>275</td>
<td>72.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60.3</td>
<td>16.0</td>
<td>0</td>
<td>264</td>
<td>76.20</td>
</tr>
<tr>
<td><strong>Olive Site</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training dataset</td>
<td>from 06:00 March 3rd to 18:00 May 9th, 2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing dataset</td>
<td>from 00:00 May 12th to 23:00 May 31st, 2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black Carbon</td>
<td>Micrograms per cubic meter</td>
<td>1620/480</td>
<td>1.1</td>
<td>0.9</td>
<td>0</td>
<td>6</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>0.8</td>
<td>0</td>
<td>4</td>
<td>0.62</td>
</tr>
<tr>
<td>Wind Direction</td>
<td>Degrees from the north</td>
<td></td>
<td>206.2</td>
<td>167.0</td>
<td>0</td>
<td>360</td>
<td>103.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>236.4</td>
<td>276.5</td>
<td>1</td>
<td>359</td>
<td>97.92</td>
</tr>
<tr>
<td>Wind Speed</td>
<td>Miles per hour</td>
<td>1620/480</td>
<td>3.5</td>
<td>3.3</td>
<td>1</td>
<td>11</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.6</td>
<td>4.3</td>
<td>1</td>
<td>12</td>
<td>2.15</td>
</tr>
<tr>
<td>Temperature</td>
<td>Degrees Fahrenheit</td>
<td></td>
<td>48.4</td>
<td>48.0</td>
<td>29</td>
<td>74</td>
<td>7.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>58.1</td>
<td>57.0</td>
<td>43</td>
<td>80</td>
<td>8.26</td>
</tr>
<tr>
<td>CAR Volume</td>
<td>Vehicles per hour</td>
<td></td>
<td>7890.0</td>
<td>9190.0</td>
<td>920</td>
<td>13714</td>
<td>3960.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7376.0</td>
<td>8278.0</td>
<td>816</td>
<td>12662</td>
<td>3740.60</td>
</tr>
<tr>
<td>TRUCK Volume</td>
<td>Vehicles per hour</td>
<td></td>
<td>81.2</td>
<td>50.0</td>
<td>0</td>
<td>366</td>
<td>76.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100.6</td>
<td>70.0</td>
<td>0</td>
<td>522</td>
<td>97.82</td>
</tr>
</tbody>
</table>

### Methodologies

*Bayesian Model Averaging for Variable Selection*

In this study the variable selection is conducted by Bayesian Model Averaging (BMA) method. BMA provides advantages over other methods on the basis of iterative significance tests, in that it accounts for the model uncertainty inherent in the variable selection processing. Hoeting *et al.*
(28) gave an in-depth review of BMA methods but generally the variable selection starts from calculating the posterior probability for model $M_k$ given the dataset $D$:

$$
\Pr(M_k | D) = \frac{\Pr(D | M_k) \Pr(M_k)}{\sum_{i=1}^{K} \Pr(D | M_i) \Pr(M_i)}
$$

(1)

where $\Pr(D | M_k) = \int \Pr(D | \theta_k, M_k) \Pr(\theta_k | M_k) d\theta_k$;

$\theta_k$ is vector of parameters in model $M_k$;

$\Pr(\theta_k | M_k)$ is prior density of $\theta_k$ under $M_k$;

$\Pr(D | M_k)$ is prior probability that $M_k$ is the true model; and it is often assumed that all models are equally likely to be chosen before observing the data.

Therefore, when comparing two possible models, by Equation (1)

$$
\frac{\Pr(M_i | D)}{\Pr(M_j | D)} = \frac{\Pr(D | M_i)}{\Pr(D | M_j)} = B_{ij}
$$

(3)

where $B_{ij}$ is the Bayes factor for $M_i$ against $M_j$. When $B_{ij} > 1$, the data favor $M_i$ over $M_j$.

Calculating the Bayes factor involves Equation (2), which is obtained by using the law of total probability by integrating over $\theta_k$. Since Equation (2) is hard to integrate exactly, the following approximation applies for large sample size:

$$
2 \log \left[ \Pr(D | M_k) \right] \approx 2 \log \left[ \Pr(D | \hat{\theta}_k, M_k) \right] - d \log (n) = -BIC
$$

(4)

where $d = \text{dim}(\theta_k)$ is the number of parameters in $\theta_k$ and $d = p + 2$ with $p$ regressors in the regression model.

Therefore the ratios represented by Bayes factors as shown in Equation (3) can be converted to differences between values of BIC. Specifically for regression, Equation (4) can be simplified as:

$$
BIC = n \log \left( 1 - R^2 \right) + p \log n
$$

(5)

where $n$ is the sample size and $R^2$ is the R-squared value from the regression.

**Model Structure Selection**

The relationship between the BC concentration (dependent variable) and regressors can be generalized as:

$$
BC_{j} \sqsubset f(X_{t,1}, \ldots, X_{t,p})
$$

(6)

Three potential model structures are identified from previous research (24 and 29) for modeling air pollutant concentrations and are tested in this study. They are expressed in Equations (7) through (9):

$$
\ln(BC_{j}) \sqsubset b_0 + b_1 \ln(X_{t,1}) + \cdots + b_p \ln(X_{t,p})
$$

(7)

$$
BC_{j} \sqsubset b_0 + b_1 X_{t,1} + \cdots + b_p X_{t,p}
$$

(8)
$\ln(BC_t) = b_0 + b_1X_{t_1} + \cdots + b_pX_{t_p}$ \hspace{1cm} (9)

**Wind Fields Incorporation**

Wind effects, in the representation of wind direction and wind speed, could generate direct impact on BC measurements. It is quite pronounced that when wind blows vehicle-emitted particulate matters from the roadway to the monitoring site, a higher concentration would be detected. Figure 2 shows the wind roses at both sites for the training dataset. The length of each vector on the wind rose graph represents the percentage of time the wind blew from each direction. Wind speed is delineated by the color and width of the brackets.

![Wind roses for the training dataset.](image)

FIGURE 2 Wind roses for the training dataset.

In the original dataset, wind direction is collected as a numeric variable. It is defined as the angle from the North to the direction where the wind originates. However, this definition poses challenges in mining the relationship between the wind direction and BC concentrations. First, when the wind direction increases from 0 to 359, it does not have a monotonic impact on air pollutant concentrations. Also, due to its circular nature, 359 degree and 1 degree have a significantly numerical difference; while they are intrinsically referring to the same northern wind. Therefore, some researchers (13 and 30) converted the wind direction ($\theta$) values to $\cos\theta$, which is denoted as Plan A in this study. Later, Arain et al. (25) proposed to combine wind direction and wind speed using the dot product operation between two vector sets to handle the wind field:

$$v_{\text{wind}} \cdot v_{\text{street}} = |v_{\text{wind}}| |v_{\text{street}}| \cos \alpha$$ \hspace{1cm} (10)
where \( \mathbf{v}_{\text{wind}} \) is the wind vector with wind direction and wind speed as its modulus; \( \mathbf{v}_{\text{street}} \) is a unit vector (with \( |\mathbf{v}_{\text{street}}| = 1 \)) with the direction from the air quality monitoring station perpendicular to the nearest roadway; and \( \alpha \) as the angle between \( \mathbf{v}_{\text{wind}} \) and \( \mathbf{v}_{\text{street}} \).

This is denoted as \textbf{Plan B} in this study. Treating wind field in this way could: (1) consolidate wind direction and wind speed into one variable; (2) denote the upwind/downwind direction by the sign of \( \cos \alpha \). For example, if \( \mathbf{v}_{\text{wind}} \) and \( \mathbf{v}_{\text{street}} \) have the same direction, \( \alpha = 0 \) and \( \cos \alpha = 1 \), which infers that the roadway is in the downwind direction (wind blows from the air quality monitoring site to roadway).

In this study, an innovative solution is proposed to transform wind direction from numerical values to categorical variables, which is denoted as \textbf{Plan C}. The full circle is firstly divided into 16 intervals with width of 22.5 degrees, and then dummy variables 1, 2, ..., 16 are assigned to the corresponding interval clockwise. For example, if wind direction is within the interval \((0, 22.5]\), the dummy variable 1 will represent this direction. Then all the possible combinations are iterated to categorize the 16 dummies into two groups while assuring that in each group selected dummies are adjacent to each other. Therefore, a grouping option of 9 and 14 means that dummies from 9 to 13 are forming one group and the rest of the dummies (from 14 to 16 and from 1 to 8) form the other group. The grouping that yields the highest \( R^2_{\text{adjusted}} \) is chosen. Then the two groups are denoted as two new dummy variables, indicating upwind/downwind directions.

\textit{Residual Analysis}

Consider the general form of regression of \( Y_t \) over \( X_{1,t}, \ldots, X_{p,t} \) is

\[
Y_t = f \left( X_{1,t}, \ldots, X_{p,t} \right) + \varepsilon_t
\]  

where the residuals \( \varepsilon_t \) contain enriched information about the regression.

In many cases residual analysis far outweighs building a prediction equation. Residual analyses are conducted by testing a series of hypotheses, including independence, normality, and constant variance and zero mean. Because leftover of BC from previous hours in the atmospheric environment can be detected in the current hour, autocorrelation problem is widely identified in short-term air quality modeling studies. Applying an Auto Regressive Integrated Moving Average (ARIMA) time series model to include the persistence information is an effective remedy for improving model accuracy.

In this study, a number of ARIMA\((p, d, q)\) models are tested on the residual time series by heuristically searching the optimal combination of \((p, d, q)\) parameters, where the differencing parameter \( d \) is determined using the KPSS test. If the null hypothesis of stationarity is accepted when the KPSS is applied to the original time series, then \( d = 0 \). Otherwise, the series is differenced until the KPSS accepts the null hypothesis. After that, \( p \) and \( q \) are selected using BIC as criteria.

Another issue is volatility clustering, where large (small) volatilities of residuals are followed by large (small) volatilities. A time series with clustered volatilities is likely to produce outlier predictions in times of high volatility, as well as heavy-tailed distributed residuals. Volatility clustering is a salient dimension that worth further investigation in nowcasting air pollutants concentrations, because if it is identified in model residuals, people should expect...
higher amount of uncertainty (wider prediction interval) when very high concentrations are predicted.

Volatility clustering is caused by a varying conditional variance, which cannot be captured by Auto Regressive (AR) models, due to its conditional homoscedasticity assumption. GARCH (General Auto Regression Conditional Heteroskedasticity) model overcomes this limitation by introducing dependent residuals to account for conditional heteroscedasticity. Specifically Equation (11) is adjusted as:

\[ Y_t = f \left( X_{t,1}, \ldots, X_{t,p} \right) + \varepsilon_t \]  

\[ \sigma_t = \sqrt{\omega + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}^2} \]  

where \( \omega, \alpha_i, \) and \( \beta_i \) are parameters. Since \( \beta_i \) brings the past values of the conditional standard deviation \( \sigma_{t-i} \) to the current one \( \sigma_t \), a larger value of \( \beta_i \) causes \( \sigma_t \) to be highly correlated with \( \sigma_{t-i} \) and gives the conditional standard deviation process a relatively long-term persistence, which indicates the duration of volatility clustering’s effects.

RESULTS AND DISCUSSION

Analyses based on collected data and proposed methodology are elaborated in this section. Before comparing different plans to incorporate the wind direction, a “baseline” model is determined by conducting preliminary variable selection and comparing three model structures. After accounting for wind direction, autocorrelation within the BC concentration time series is resolved by introducing an AR(1) component, to conclude a final model. Residuals are further analyzed using GARCH techniques, bringing insights to the application of the final model for nowcasting purposes. Finally, the predictive power of the final model is tested against several commonly-used measures.

Preliminary Variable Selection and Model Structure Determination

Variable selection is critical to both model fitting and the interpretative power of proposed models. A preliminary variable selection is conducted over three potential model structures using BMA methods to identify significant regressors. The results are shown in Figure 3 for the Duwamish site and Figure 4 for the Olive site.
FIGURE 3 First round variable selection for the Duwamish site.

From left to right, the three graphs in each figure correspond to Equation (8) “no transform”, Equation (9) “only BC transformed”, and Equation (7) “all transformed”, respectively. A red (blue) stripe in the figures indicates a positive (negative) correlation, while a beige stripe indicates that the corresponding regressor is not significant. For each model structure, BMA evaluates several candidate models with different sets of regressors and lists all the models based on their posterior probability. The two figures indicated that for the Duwamish site, the significant explanatory variables include TRUCK, TEMP, and WIND SPEED (temperature is significant in two out of three models and therefore still considered in this preliminary round);
while for the Olive site, the subset comprises of CAR, TRUCK and WIND SPEED. Vehicle speed at the Duwamish site is not consistently significant across all candidate models. This may result from the nature of speed data at the study site: vehicle speeds on SR-99 do not vary significantly (with standard deviation of 4.77 mph or 7.67 km/h). Also wind direction is not included in this preliminary round and left for further discussion.

The selected regressors for each site from the preliminary round are then applied to determine the model structure. Note that all three model structures include intercepts to account for the miscellaneous contributors to near-road BC concentrations, such as urban background concentration. The $R^2_{adj}$ values for three model structures are listed in Table 2. It is indicated that for both sites, applying log-transform to the BC concentration (Equation 9) yields best results. Therefore this model structure is chosen in the following discussion, to further incorporate the wind field effect.

**TABLE 2 Model Structure Selection**

<table>
<thead>
<tr>
<th>Model Structure</th>
<th>Adjusted R-squared value</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Duwamish</td>
<td>Olive</td>
</tr>
<tr>
<td>No transform</td>
<td>0.3288</td>
<td>0.3537</td>
</tr>
<tr>
<td>Only BC transformed</td>
<td>0.4204</td>
<td>0.3768</td>
</tr>
<tr>
<td>All transformed</td>
<td>0.4152</td>
<td>0.3507</td>
</tr>
</tbody>
</table>

**Incorporating Wind Field**

Using the methodology described in the previous section, Plans A through C are applied to the training dataset for the wind field processing. The iteration results for Plan C group intervals from 9 to 13 together (coded as zero in the Wind Dummy) for the Duwamish site, and group intervals from 5 to 8 together (coded as zero in the Wind Dummy) for the Olive site. All three options are incorporated into the preliminary models and comparison is shown in Table 3:

**TABLE 3 Comparison across Three Plans of Treating Wind Field**

<table>
<thead>
<tr>
<th>Site</th>
<th>Duwamish</th>
<th>Olive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Structure</td>
<td>ln(BC) = $b_0 + b_1$TRUCK + $b_2$TEMP + $b_3$WS</td>
<td>ln(BC) = $b_0 + b_1$TRUCK + $b_2$CAR + $b_3$WS</td>
</tr>
<tr>
<td>Wind Direction</td>
<td>$R^2_{adj}$</td>
<td></td>
</tr>
<tr>
<td>Preliminary Model</td>
<td>0.4204</td>
<td>0.3768</td>
</tr>
<tr>
<td>Plan A</td>
<td>0.4438</td>
<td>0.3913</td>
</tr>
<tr>
<td>Plan B</td>
<td>0.2907</td>
<td>0.3774</td>
</tr>
<tr>
<td>Plan C</td>
<td>0.5058</td>
<td>0.4230</td>
</tr>
</tbody>
</table>

Both Plan A and Plan B essentially transform the wind direction using trigonometric functions, to resolve the inconsistency between the recorded numerical values and the actual wind effect. However, even after the transformation, neither Plan A nor Plan B could indicate a real-world upwind/downwind scenario. By contrast, Plan C not only yields the highest $R^2_{adj}$,
but also explicitly highlights upwind/downwind directions. The parameter estimation for the selected model also validates this point. For the Duwamish site, if the wind is coming from the Puget Sound (Zones 9 to 13), the BC concentration is reduced. For the Olive site, if the wind blows from the I-5 freeway, the BC concentration will be higher. This consistency in modeling results for the wind field reflects the validity of Plan C.

**Final Model**

The Durbin-Watson statistics are 0.647 and 0.597 for Duwamish and Olive sites respectively, both with zero p-values, which indicates a strong dependency in the residual time series. Using the approach proposed in the previous section, an AR(1) structure is suggested for the residual time series. The final model is therefore an AR(1) univariate BC model with a series of exogenous regressors. Parameter estimation is summarized in Table 4, where WDC1 refers to the wind direction.

| TABLE 4 Parameter Estimation of the Final Model | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------------------|----------------|-------------|----------|----------|
| Duwamish                 |              |             |          |          |
| (Intercept)              | -0.0754      | 0.02658     | -2.84    | 0.005    |
| Lag_BC                   | 0.6872       | 0.01345     | 51.10    | 0.000    |
| TRUCK                    | 0.0023       | 0.00017     | 13.77    | 0.000    |
| Wind_Speed               | -0.0664      | 0.00425     | -15.61   | 0.000    |
| WDC1                     | 0.2533       | 0.02136     | 11.86    | 0.000    |
| Adjusted R-squared       | 0.7680       |             |          |          |
| Olive                    |              |             |          |          |
| (Intercept)              | 0.0993       | 0.02289     | 4.34     | 0.000    |
| Lag_BC                   | 0.7029       | 0.01441     | 48.77    | 0.000    |
| TRUCK                    | 0.0036       | 0.00023     | 15.46    | 0.000    |
| Wind_Speed               | -0.0630      | 0.00572     | -11.02   | 0.000    |
| WDC1                     | -0.1032      | 0.01590     | -6.49    | 0.000    |
| Adjusted R-squared       | 0.7520       |             |          |          |

With the comparatively concise model structure (only four regressors), the models yield a good estimation comparable to those reported in the previous research. Introducing the AR(1) removes the CAR variable from the Olive site model, and the Temperature variable from the Duwamish site model. Compared to other similar studies, the absence of Temperature might be attributable to the fact that the dataset used in this study were collected in spring and early summer, when temperature variation was not significant in Seattle. The same set of regressors for both sites reflects the very similar correlation between near-road BC concentration and impact factors. The significance of Lag_BC (last hour’s BC concentration) indicates a strong air pollutant residue in a near-road context. Also by comparing these two sites, it is found that the Olive site is associated with a stronger “residue” than the Duwamish site, mainly due to the “street canyon effect” in the Seattle downtown area. Wind field, including both the wind speed and wind direction, is major impact factor in determining short-term BC concentrations. Higher wind speed would lower the concentration as it accelerates the dispersion of BC. The sign of WDC1 is consistent with the actual spatial relationship between the roadways and the air quality.
monitoring sites. It can be concluded from both models that for near-road BC concentration, TRUCK volume is the only traffic-related contributor and any additional truck will result in 0.23% and 0.36% increases of BC concentration for Duwamish and Olive sites, respectively. The smaller coefficient for TRUCK at the Duwamish site to certain extent reveals the effectiveness of the Port of Seattle Clean Truck Program, which was enacted in 2008 requiring drayage trucks serving the ports to be equipped with 1994 or newer engines. Therefore although at the Duwamish site there is higher truck volume percentage, the environmental impact (speaking of near-road BC concentration) at this site is lower than its downtown peer, due to a cleaner truck fleet from the port regulation.

**Residual Analysis: GARCH Models**

Before applying the final models to nowcasting, special attention should be paid to the volatility clustering and heavy-tailed distribution of residuals, as they influence the prediction interval. The Jarque-Bera test of normality strongly rejected the null hypothesis that the distribution of residuals is normal ($p$ value is zero). It further demonstrates from the QQ plot a heavy tail effect on both ends in the residual distribution. It is suggested that t-distribution with $df=5.4$ is more suitable in describing the residuals. A heavy-tailed distribution as the t-distribution here implies that the regression model is more prone to producing more outliers and we need to pay attention to very high (or low) predictions. The residuals are further analyzed using a GARCH(1,1) model and the parameter estimation is summarized in Table 5.

**TABLE 5 Parameter Estimation for GARCH Models**

|          | Estimate | Std. Error | t value | Pr(>|t|) |
|----------|----------|------------|---------|----------|
| **Duwamish** |          |            |         |          |
| mu       | 0.0000   | 0.00891    | 0.00    | 1.000    |
| omega    | 0.0773   | 0.02815    | 2.75    | 0.006    |
| alpha1   | 0.1170   | 0.03239    | 3.60    | 0.000    |
| beta1    | 0.5080   | 0.15414    | 3.30    | 0.001    |
| shape    | 6.1200   | 0.73375    | 8.35    | 0.000    |
| **Olive** |          |            |         |          |
| mu       | 0.0000   | 0.00732    | 0.00    | 1.000    |
| omega    | 0.0451   | 0.01635    | 2.76    | 0.006    |
| alpha1   | 0.1090   | 0.03388    | 3.22    | 0.001    |
| beta1    | 0.4150   | 0.18429    | 2.25    | 0.024    |
| shape    | 8.8700   | 1.70001    | 5.22    | 0.000    |

For both sites, $\alpha_i$ indicates a significant volatility clustering within the residuals. The relatively large values of $\beta_i$ (0.51 for the Duwamish site, and 0.42 for Olive) cause $\sigma_i$ to be highly correlated with $\sigma_{i-1}$ and gives the conditional standard deviation process a relatively long-term persistence (by Equations 13a and 13b). The shape parameter being significant again confirms the good fit of a t-distribution for the residuals rather than a normal distribution.

Using the GARCH model for the residuals would provide more accurate prediction intervals if the model is to be used for nowcasting. Because of the volatility clustering effect, people should expect wider prediction intervals to account for the higher amount of uncertainty.
during higher volatility periods. Similarly, the prediction intervals will narrow in times of lower
volatility. Prediction intervals using an ARMA model without conditional heteroscedasticity
cannot adapt in this way. The persistence detected by larger $\beta$ value also offers insights that
once the predictions are with high volatility, it will last for a longer period; therefore for the next
a few hours the predictions will be with high volatility as well.

Predictions

The fitted regression models are then applied to the testing datasets. The ability of the models to
produce accurate predictions is judged against the following statistical performance metrics:
Root Mean Square Error (RMSE), Normalized RMS (NRMS), Mean Absolute Percentage Error
(MAPE), Index of Agreement (IA), Fractional Bias (FB), and Prediction Accuracy (PA), where $n$
is the number of observations, $P_i$ is the predicted value, $O_i$ is the observed value, $\bar{O}$ is the mean
of observed values, $S_p$ is the standard deviation of predicted values, and $S_o$ is the standard
deviation of observed values. The results are summarized in Table 6.

\begin{equation}
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (O_i - P_i)^2}
\end{equation}

\begin{equation}
NRMS = \frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} (\bar{O} - O_i)^2}
\end{equation}

\begin{equation}
IA = 1 - \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (|P_i - \bar{O}| + |O_i - \bar{O}|)^2}
\end{equation}

\begin{equation}
FB = \frac{(\bar{O} - \bar{P})}{0.5 \times (\bar{O} + \bar{P})}
\end{equation}

\begin{equation}
PA = \frac{\sum_{i=1}^{n} [(P_i - \bar{P})(O_i - \bar{O})]}{(n-1)S_pS_o}
\end{equation}

<table>
<thead>
<tr>
<th>TABLE 6 Prediction Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>Normalized RMS</td>
</tr>
<tr>
<td>Index of Agreement</td>
</tr>
<tr>
<td>Fractional Bias</td>
</tr>
<tr>
<td>PA</td>
</tr>
</tbody>
</table>
Table 6 shows that with only four regressors, the final models (as shown in Table 4) present good agreements (IA values 0.94 and 0.892). The Duwamish site (Industrial/Arterial) obtains better agreement than the Olive site (Freeway). As diesel exhaust is a major source to BC concentrations during spring/summer season and as both models suggest, only TRUCK volume matters in the relationship and the higher truck percentage at the Duwamish site should account for the better agreement. For real-world applications, truck volume inferred from historical archive and daily patterns can be used as model input, while meteorological conditions can be retrieved from the weather nowcast.

CONCLUSIONS

A plethora of existing research has identified high air pollutant concentrations in close vicinity of major roadways due to vehicle emissions. A coherent modeling procedure is demonstrated using two urban sites with heavy traffic in Seattle, U.S. to capture the relationship between near-road hourly BC concentrations and on-road traffic and meteorological conditions. The motivation of building up a short-term linear model is to facilitate the nowcasting of near-road air pollution for human-exposure assessment and public information access.

The final model has a concise model structure with log-transformed BC concentration as the dependent variable and only four regressors: the past-hour BC concentration, TRUCK volume, wind speed and wind direction. BMA techniques are used for variable selection. By taking the inherent model uncertainty in the variable selection into account, BMA averages over the best models according to approximate posterior model probability. Using vehicle classification data, impacts from passenger cars are found trivial and insignificant and the final models reveal the strong bond between BC and truck volume. It implies that to reduce the near-road Black Carbon concentrations, improved management of freight activities is the key. A novel approach, based on iterating all the possible groupings of 16 sub-directions of wind, is deployed to generate one wind direction dummy (WDC1). This approach greatly enhances the model performance. It also identified the upwind/downwind directions mathematically rather than arbitrarily.

Models for both sites achieved satisfying result in model fitting and predictive accuracy. $R^2_{adj}$ are 0.7680 and 0.7520 for the Duwamish site and the Olive site respectively. IAs are 0.94 and 0.892 respectively, and PAs are 0.901 and 0.813, suggesting good agreement between model predictions and actual observations. This good agreement is partially contributable to the massive traffic volume on site, which is a major contributor of BC concentration, yet more importantly it is attributed to the “fine-tuned” modeling approach as demonstrated in this study.

Residual analyses are thoroughly conducted and reveal certain intriguing characteristics about the model residuals. As the autocorrelation issue is mitigated by the introduction of AR(1) component (the past-hour BC concentration), the volatility clustering and the heavy-tailed distribution of residuals are successfully identified and modeled by the General Auto Regressive Conditional Heteroscedasticity (GARCH) model. This brings valuable insights to the prediction intervals and outliers. A t-distribution is found to be more suitable for model residuals and implies that the model is prone to producing more outliers. Also wider prediction intervals should be expected in times of higher volatility. And for both datasets studied, longer persistence of volatility clustering is detected.
Some insights from this study include: first, a structurally concise and computationally efficient model with satisfactory predictions; second, a coherent methodological approach for modeling and nowcasting hourly near-road BC concentrations, including variable selection, model structure, and residual analysis; third, an innovative solution to processing the wind direction for future related research; last but not least, the model highlights the imperative of managing truck fleet to mitigate BC impacts. It also indirectly reflects the effectiveness of freight regulation for better near-road environment.

Future work involves comparing the proposed model with other methodologies as well as modeling other near-road air pollutants. In either of the effort, the procedure elaborated in this study, such as variable selection, wind field incorporation, and residual analyses, provides pertinent references. Additionally it is worth endeavoring to integrate the proposed model with traffic archived data and routine weather forecast models, in order to enhance the model’s operational feasibility of nowcasting.

REFERENCES


