Understanding the PM$_{2.5}$ imbalance between a far and near-road location: Results of high temporal frequency source apportionment and parameterization of black carbon

U.M. Sofowote$^{a,*}$, R.M. Healy$^a$, Y. Su$^a$, J. Debosz$^a$, M. Noble$^a$, A. Munoz$^a$, C.-H. Jeong$^b$, J.M. Wang$^b$, N. Hilker$^b$, G.J. Evans$^b$, P.K. Hopke$^{c,d}$

$^a$ Environmental Monitoring and Reporting Branch, Ontario Ministry of the Environment and Climate Change, Toronto, Canada
$^b$ Southern Ontario Centre for Atmospheric Aerosol Research, University of Toronto, Toronto, Canada
$^c$ Center for Air Resources Engineering and Science, Clarkson University, Potsdam, NY, USA
$^d$ Department of Public Health Sciences, University of Rochester School of Medicine and Dentistry, Rochester, NY, USA

**A R T I C L E   I N F O**

Keywords:
- Multilinear engine (ME)
- Multi-time resolution
- Receptor modelling
- PM$_{2.5}$
- Black carbon
- Aerosol
- Nonlinear modelling
- Machine learning
- Source apportionment

**A B S T R A C T**

The differences in PM$_{2.5}$ concentrations between two relatively close stations, one situated near a major highway and the other much more distant were used to develop a protocol for determining the impact of highway traffic on particulate matter concentrations at the roadside. The roadside station was $<15$ m away from the edge of a major highway while the other was located $\sim 170$ m away. The roadside station contains a suite of continuous instrumentation capable of near-real-time speciation of PM$_{2.5}$. The particulate matter difference, formally termed the PM$_{2.5}$ imbalance was arbitrarily defined as a case wherein $|\text{Near-road PM}_{2.5} - \text{Far from road PM}_{2.5}|/\text{Near-road PM}_{2.5} \geq 50\%$. Of interest was the variation of multi-time factors based on ME2 analyses of the speciation data from the roadside station during these imbalance events. Of the 7 mass-contributing ME2 factors, a black carbon factor was determined to be the major cause of the PM$_{2.5}$ imbalance and was especially dominant for the case when PM$_{2.5}$ concentrations at the roadside station were greater than the farther-station PM$_{2.5}$. The black carbon concentrations observed during these specific events were further regressed against other traffic-related and meteorological parameters with two nonlinear optimization algorithms (generalized reduced gradient and rules ensemble) in our attempts to model any potential relationships. It was observed that the traffic counts of heavy duty vehicles (predominantly diesel-powered) dominated the relationship with black carbon while contributions from light duty vehicles were negligible during these [PM$_{2.5}$]Roadside $> \text{[PM}_2.5\text{]}$Further events at the roadside station. This work details the most critical ways that highway traffic can contribute to local ambient PM$_{2.5}$ concentrations that commuters are exposed to and will be important in informing policies and strategies for particulate matter pollution reduction.

**1. Introduction**

Exposure to PM$_{2.5}$ can potentially be deleterious to health (Pope et al., 2004; Thurston et al., 2016), and remains an issue of active research in this region (Requia et al., 2017). While PM$_{2.5}$ can have local primary sources, it has been observed that across most locations in the Golden Horseshoe region of southern Ontario, the annual average concentrations determined at the air quality monitoring stations operated by the Ministry of the Environment and Climate Change (MOECC) in general do not differ by more than a few micrograms per cubic metre (Ontario Ministry of the Environment and Climate Change, 2017). This result indicates that PM$_{2.5}$ is largely a regional pollutant comprised mainly of secondary products. Thus, to a first approximation, PM$_{2.5}$ concentrations at locations that are separated by relatively short distances can be expected to be very similar. However, it is important to study the impact of local activities on this assumption. Recently, there has been increased interest in the effects of vehicular traffic on pollutants in the near-road environment. To assess the effect of a major highway on PM$_{2.5}$, the concept of PM$_{2.5}$ imbalance is introduced. The imbalance is defined as the difference in PM$_{2.5}$ concentrations between two nearby stations that are otherwise similar except for the proximity of one of them to an adjacent source such as a major roadway. Thus, both stations may be assumed to be affected by urban and regional PM$_{2.5}$ but the one closer to the roadway may be more heavily

*$^*$ Corresponding author.

E-mail address: Uwayemi.Sofowote@ontario.ca (U.M. Sofowote).

http://dx.doi.org/10.1016/j.atmosenv.2017.10.063
Received 6 August 2017; Received in revised form 21 October 2017; Accepted 29 October 2017
Available online 01 November 2017
1352-2310/ Crown Copyright © 2017 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).
influenced by the vehicular traffic. The objective is to determine if indeed, vehicles on the adjacent roadway produce sufficient emissions of traffic-related air pollutants (TRAPs) that can cause significant differences in PM$_{2.5}$ composition and concentration.

To get a clear understanding of what species vary due to an adjacent pollution source such as local traffic especially to design control strategies, pollutant gradients studies can be performed. Specifically, TRAPs of interest are simultaneously determined at two (or more) locations differing in their distance from a major roadway. This approach has previously been applied to volatile organic compounds VOCs (e.g., Olson et al., 2009) and ultrafine particles UFP (e.g., Jeong et al., 2015) albeit for relatively short periods of time. For these homologous types of pollutants, the approach is straightforward since only a single instrument type need be replicated and deployed. However, for PM$_{2.5}$ composition, this approach is almost prohibitively challenging because PM$_{2.5}$ is a mixture of different components/species that need to be resolved continuously to determine the chemical identities of the contributory sources. If a TRAP gradient were to be attempted in the traditional way, various monitoring instruments and systems would have to be exactly replicated at all chosen sampling locations. Given the insurmountable constraints of costs, available space and time, we have attempted an innovative way to employ the continuous TRAP data from the near-road location alone coupled with factor analysis on this data in the form of high-time resolution (Ogulei et al., 2005; Zhou et al., 2004) multilinear engine (ME2; Paatero, 1999) analysis and the bulk PM$_{2.5}$ mass differences between the two monitoring locations to reveal the underlying causes of PM$_{2.5}$ imbalance in the form of factors and contributions when present. The overarching aim is to determine whether such factors that are responsible for the PM$_{2.5}$ imbalance are directly linkable to vehicular traffic characteristics.

2. Methods

2.1. Sampling sites and data description

The MOECC in conjunction with partners at Environment and Climate Change Canada (ECC) and the Southern Ontario Centre for Atmospheric Aerosol Research (SOCAAR) at University of Toronto operates a unique research station that is < 15 m from the edge of Highway 401 between the Islington Avenue and Weston Road overpasses. This section of the highway contains 18 lanes of fast-moving east and westbound traffic and experiences an annual average daily traffic volume of 405,500 vehicles (Ministry of Transportation of Ontario, 2013). At this station, instrumentation capable of continuous measurements of gaseous and particulate species were fitted with PM$_{2.5}$ inlets at a sampling height of roughly 6 m a.g.l. and operated to generate hourly and sub-hourly concentrations. Thus, PM$_{2.5}$ mass, speciation components – organic ions, sulphate, nitrate and ammonium ions, black carbon (BC) and metals and gaseous pollutants like O$_3$, NO$_x$, SO$_2$, CO and CO$_2$ as well as meteorological parameters (Vaisala, 2012) i.e., wind direction (WD), wind speed (WS), relative humidity (RH) and ambient temperature (ATEM) were determined and automatically data-logged on a semi-continuous basis. The table below (Table 1) shows the complete list of chemical species determined continuously and used in this work. Principles of operation and basis for use of these instruments have been discussed in previous work and other reports.

The ACSM was retro-fitted with a PM$_{2.5}$ lens and a new capture vaporizer (Xu et al., 2017) for the majority of the study period, i.e., from February 2016 but no major differences were observed when compared with the previous PM$_{1}$ lens data, likely due to the relatively fresh nature of the bulk of the organic aerosols being sampled at this location. Raw ion signals were converted to mass spectra using the ACSM data acquisition (DAQ) software with the embedded fragmentation table. An additional module (ACSM Error Calc) was subsequently used to generate uncertainty values for all measured m/z species. The data used in this study was collected between June 18, 2015 and November 22, 2016 with a few periods of one of more instruments not being operational. Traffic count data was recorded with a Wavetronix (2008) SS125 HD sensor and 1-h data were validated for a period of 150 days between March 15 and November 22, 2016.

Located to the south of this roadside station is the regular air quality (AQ) monitoring station referred to here as the Upper Garage (U.G./Station 35125) site also operated by the MOECC where only criteria air pollutants (e.g., O$_3$, NO$_x$, and PM$_{2.5}$) are continuously monitored with similar instruments (for this work, the relevant species, PM$_{2.5}$ is monitored with a SHARP 5030) and publicly reported on an hourly basis as part of the Ministry’s mandate to provide air quality information to Ontarians. The shortest distance between this regular AQ station and the edge of Highway 401 is 167 m (343.59°). Although this station is labelled a roadside station in the provincial air quality reports (Ontario Ministry of the Environment and Climate Change, 2017), it is best regarded as a far-roadside station in comparison with the Highway 401 station that can be designated as a near-roadside station. The figure below (Fig. 1; note imbalance pollution rose adjacent to each site to be discussed later) shows the distance and locations of both stations in relation to each other and to the highway.

2.2. Model description

Multi-time factor analysis using ME2 has been described in detail previously (Ogulei et al., 2005; Zhou et al., 2004). In our implementation, ME2 was applied with the aim of preserving the high frequency of any potential roadside-derived factors while achieving mass closure on sampled PM$_{2.5}$ at the roadside station. This station, being close to a highly dynamical potential source required the use of data in their native time resolutions so as to maximally retain the information contained from this source. ME2 allows the fitting of the basic receptor modelling equation:

$$x_j = \sum_{k=1}^{K} \beta_k f_{kj} + e_j$$

Table 1

<table>
<thead>
<tr>
<th>Continuous Instrument</th>
<th>Full name and manufacturer</th>
<th>Variables/species</th>
<th>Sampling duration used (Hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACSM</td>
<td>Aerosol chemical speciation monitor (Aerodyne Research Inc.)</td>
<td>Organic ions m/z 12–130 (+mode), NH$_4^+$, Cl$^-$, NO$_3^-$, SO$_2^-$, Si, K, Ca, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, As, Se, Br, Rh, Ag, Cd, Ba, Hg, Pb</td>
<td>0.5</td>
</tr>
<tr>
<td>XACT 625</td>
<td>XACT 625 Ambient metals/fence line monitor (Cooper Environmental Services)</td>
<td>BC6 (880 nm)</td>
<td>0.5</td>
</tr>
<tr>
<td>AE33</td>
<td>Aethalometer AE33 (Magee Scientific)</td>
<td>BrCl = BC1 (370 nm) – BC3 (520 nm), BrC2 = BC3 – BC6</td>
<td>0.5</td>
</tr>
<tr>
<td>SHARP 5030</td>
<td>Synchronized Hybrid Ambient Real-time Particulate Monitor; SHARP 5030 (Thermo Fisher Scientific)</td>
<td>PM$_{2.5}$</td>
<td>0.5</td>
</tr>
</tbody>
</table>
where \( x_i \) is the measured value, \( e_i \) is the residual (unfitted) value, \( f_{ij} \) and \( g_{sk} \) are elements of the factor profile F and contribution G matrices, respectively. The prior equation can be flexibly expanded to account for different time-bases as:

\[
x_{sk} = \frac{1}{t_{s2} - t_{s1} + 1} \sum_{i=1}^{P} f_{ij} \sum_{k=1}^{G} g_{sk} f_{jk} + e_{sk}
\]

(2)

where \( x_{sk} \) is the concentration of the \( j \)th species in the \( s \)th sample, \( t_{s1} \) and \( t_{s2} \) represent the start and end times for the \( s \)th sample in discrete time units that are integer multiples of the smallest sampling interval (30 min in this work), \( g_{sk} \) is the contribution of source \( k \) during the time duration of the \( s \)th sample; \( e_{sk} \) is the residuals of \( x_{sk} \); \( \eta_j \) are adjustment factors for replicated species measured by different samplers (no species was replicated in this work, so \( \eta_j = 1, \forall \ x_{sk} \)). If \( t_{s1} \) and \( t_{s2} \) are the same, Eq. (2) will collapse to Eq. (1).

The smoothing equation:

\[
g_{sk+1,k} - g_{sk,k} = 0 + \varepsilon_i
\]

(3)

is also included in the fitted model to minimize unrealistic values of \( \varepsilon_i \) using least squares, since, as explained by Ogulei et al. (2005), a source may contain only species measured at a relatively lower time resolution or both low and high time resolved species. When a source contains some high-time-resolution species, reducing \( \varepsilon_i \) leads to an increase in \( e_{sk} \) and the high temporal variations tend to be conserved (Ogulei et al., 2005; Zhou et al., 2004).

To obtain a solution for the receptor modelling problem, ME2 requires that measurement uncertainties \( u_{ij} \) associated with the determination of \( x_{ij} \) be included. This \( u_{ij} \) is solved using a suitable ME2 error model; in this case:

\[
u_{ij} = c_1 + c_2 \max(\|\varepsilon_{ij}\|, \|\eta_{ij}\|)
\]

(4)

where, \( c_1 \) is a multiplier for adding extra uncertainty, \( x_{ij} \) is the observed value, \( y_{ij} \) is the modelled value of \( x_{ij} \) and \( c_2 \) is the measurement or estimated error that was assigned as discussed for each instrument case below. When MDL values were available, they were used as the \( c_1 \) or the basis for optimizing it. The SHARP 5030 gives the bulk PM\(2.5\) mass concentration that is included in the ME2 analysis as the sum of species. It has a reported minimum detectable limit of 0.5 \(\mu g/m^2\) at one hour resolution (Thermo Fisher Scientific Inc, 2007). This value was rounded to 1 \(\mu g/m^2\) for 30-min data. However, this variable, being the sum of species, had an uncertainty scaled up by a factor of 1000 in the actual analysis.

The aethalometer (BC) has a sensitivity of 0.03 \(\mu g/m^3\) at 1 min resolution and 1-hr detection limit at \(0.005 \mu g/m^3\) (Magee Scientific, 2015b). For 30 min, this was estimated at \(0.006 \mu g/m^3\). When two black carbon channels were subtracted (i.e., for BrC1 and BrC2; see Table 1 for wavelengths of aethalometer channels used), the resultant uncertainty was computed as (\(\sqrt{2}\)) \times 0.006 \(\mu g/m^2\). The mass absorption cross-section (MAC) at 880 nm for this aethalometer was determined to be 11.45 \(m^2g^{-1}\) (Healy et al., 2017). This value was used to convert the absorption of BC rather than the default value (7.77 \(m^2g^{-1}\)). The reported concentrations of BrC1 and BrC2 are thus only rough estimates of ultraviolet-absorbing PM or brown carbon assuming they had the same MAC values as BC.

The 1-hr MDLs for the Xact 625 were optimized and used as the basis for \( c_1 \). The procedure for this optimization involved computing 5- and 7-factor solutions (FS) ME2 on a \(\sim25\%\)-subset of the metals data only (\(\sim2000\) hourly samples) with MDL as \( c_1 \); then generating histograms of the scaled residuals to determine the shape and symmetry of their distributions (Zhou et al., 2004). Ideally, the scaled residuals (x) should be normally or near-normally distributed, say, in discrete bins such as \(-1 \leq x < 0, 0 \leq x \leq 1, > 1\). For most metals in this work, the MDL was used as the \( c_1 \) because the scaled residuals were fairly symmetrical, however, for K, Cr, V, As, Se, Ba and Pb, slight adjustments to the MDLs were necessary to make the distribution more symmetrical (see Table S1 and Fig. S1 for before and after \( c_1 \) and scaled residual distributions).

As discussed earlier, the ACSM Error Calc module was used to generate the ion uncertainties. Error Calc assumes the ion counts are in a Poisson distribution (see details in the supplementary information). This distribution is then used to estimate the uncertainties associated with the determination of the ion signal.

In summary, for the complete ME2 input data matrix, there were 24,742 discrete time units, 31,119 samples, 133 variables and \( c_2 \) was set at 0.1. The \( c_1 \) and \( c_2 \) values for the smoothing equations were set at 1.5 and 0.01 respectively, ME2 error model code \(-14\) was used and 15 random runs were initiated and between-run variations in Q were monitored. Missing values were treated in a hierarchy: if the ACSM data was missing for a given sampling interval, the data was completely removed but if the metals were missing, the ACSM data was kept while the metals data was replaced with the missing value indicator ~999. If single time-samples were missing both the ACSM and metals data, these missing data values were replaced with ~999. If blocks of time-samples were missing both ACSM and metals data, the time blocks were completely removed from the data set. Data substitutions were kept to a minimum and most data was used ‘as is’. However, if the data point was negative and very small numerically, i.e., \(\sim-2.5 \times c_1(y_{ij})\), it was replaced with \(\sim2.5 \times c_1(y_{ij})\) so that the model does not attempt to fit these values. Ions \(m/z\) 12 and 35 were downweighted by scaling up their uncertainties ten-fold due to their low abundances.

3. Results and discussion

3.1. Temporal trends of PM\(2.5\) imbalance

During this period of study, it was observed that PM\(2.5\) generally showed little difference in value between the sampling locations. The
bulk PM$_{2.5}$ mass concentrations at the U.G. and H401 were generally ‘equivalent’ for the entire ~18 months of study suggesting that this pollutant is indeed of urban and regional origins. For valid data, equivalence is arbitrarily defined as an absence of imbalance of PM$_{2.5}$ at the two sites. The working ‘imbalance’ criterion in turn is defined as:

\[
\frac{[\text{PM}_{2.5}]_{\text{H401}} - [\text{PM}_{2.5}]_{\text{U.G.}}}{[\text{PM}_{2.5}]_{\text{H401}}} \geq 50\%
\]  

(5)

so that practically, any value that could be rounded to \(\pm 50\%\), i.e., 

\(-54.4\% \leq \frac{[\text{PM}_{2.5}]_{\text{H401}} - [\text{PM}_{2.5}]_{\text{U.G.}}}{[\text{PM}_{2.5}]_{\text{H401}}} \leq 54.4\%\)

is not regarded as reflecting an imbalance. This arbitrary level is chosen as a balance between sensitivity and precision between two SHARP instruments. Note that the imbalances of the few PM$_{2.5}$ values between 

\(-3\ \mu g/m^3 \leq [\text{PM}_{2.5}] < 3\ \mu g/m^3\) at either station were also down-weighted by 0.25, thus making them equivalent. PM$_{2.5}$ values less than 

\(-3\ \mu g/m^3\) were invalidated.

The findings indicated that out of approximately 23 000 valid half-hourly data, only about 15% showed an imbalance. Thus, the two stations were equivalent with regards to PM$_{2.5}$ ~85% of the time. However, for the approximately 3550 non-equivalent half-hourly data points, 73% showed $[\text{PM}_{2.5}]_{\text{H401}} > [\text{PM}_{2.5}]_{\text{U.G.}}$, the remainder were $[\text{PM}_{2.5}]_{\text{U.G.}} > [\text{PM}_{2.5}]_{\text{H401}}$. Fig. 2 below shows that the imbalance appeared to have very distinct temporal trends.

Diurnal hourly (Fig. 2a) and day of week (2b) patterns seem to indicate weekday rush hour traffic as the likely cause for the H401 > U.G. It is hypothesized that since the U.G. is relatively closer to urban-residential areas, highly localized activity in those areas may influence the U.G. more with the urban/regional PM$_{2.5}$ especially on weekends and after-workhour periods, thus, may be the cause of U.G. > H401 PM. Fig. 2a indicates that while the H401 > U.G. PM imbalance events frequency was always greater than the U.G. > H401 PM imbalance, the greatest frequency occurred during the morning rush hour period (typically 6:00–8:00 a.m.). In fact, the average ratio of the imbalances between 5:00–7:00 a.m. is 7.3:1 whereas, the maximum ratio at any other period is 4.7:1 (9:00 p.m.) indicating that there is a lesser likelihood that the U.G. > H401 PM imbalances will occur during the morning rush hour. Sorting the imbalances by wind speeds (Fig. 2c) generally indicates that they occurred when wind speeds were 1–2 m/s, i.e., low while wind directions indicate the highway as the major source of the imbalance at the H401 station (see Fig. 1 for a polar plot of histogram in Fig. 2d).

3.2. Multi-time ME2 factors and contributions

Given the temporal characteristics of the imbalances, additional evidence in the form of chemical fingerprints of contributory factors to the observed PM$_{2.5}$ can help to confirm that vehicular traffic emissions from the highway are the cause of the imbalance. Therefore, a multi-timescale ME2 analysis was performed on the entire data set wherein sampling data at their native time resolution, (i.e., without averaging to a larger time-base) was used to preserve information from the fast-moving traffic on the highway. The post-ME2 analysis extracted the specific dates and times previously associated with the PM$_{2.5}$ imbalances to monitor any specific factors whose values exceed their typical concentrations.

For the multi-time factor analyses, ME2 was used to extract 7–9 factor solutions (FS). An 8FS was chosen since it gave the best balance of minimal variations of Q (the minimized objective in the factor analysis) among all 15 runs upon convergence, reasonable distributions of scaled residuals and factor splits. While 8 factors were computed, only 7 factors contributed to PM$_{2.5}$ mass (Fig. 3 and last row of raw F matrix in Table S2). In all, these factors combined for a slope of 0.900 and an $R^2$ of 0.675 when compared with the observed SHARP PM$_{2.5}$ mass (Fig. S2).

Several BC-rich factors were found. The first was almost entirely dominated by BC measured at 880 nm i.e., BC6, (> 90% by mass; Fig. 3a). On average, this BC-rich factor, contributed about 23% to the
Fig. 3. Plots of scaled BC, BC-BrC, RD, pNO3, Acid-A, OOA, and pSO4 ME2 F (a, c, e, i, k, m; ng/ng) and G (b, d, f, h, j, l, n; ng/m³) matrices as seasonal variations of mass contributing factors resolved between weekday and weekends (box = IQR, whiskers = 10th and 90th percentiles, horizontal line = median, and asterisk = mean). Horizontal axes labels on left are m/z 12-130, NH₄⁺, NO₃⁻, SO₄²⁻, Cl⁻ ions, BC6, BrC1, BrC2, and metallic elements Si, K, Ca, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, As, Se, Br, Rb, Ag, Cd, Ba, Hg, and Pb.
total PM$_{2.5}$ mass during the period of study. Willis et al. (2016) observed a similar BC-rich factor in downtown Toronto with values larger near a busy roadway. The temporal trend in Fig. 3b shows that this factor was most abundant in the summer and minimal in the winter. This difference is hypothesized to be related to the seasonal fuel composition change especially in diesel fuel oil used by heavy duty vehicles (Agency for Toxic Substances and Disease Registry, 1995; Chevron Global Marketing, 2007), although seasonal changes in fuel composition in this region, wherein a slight wintertime reduction in vehicle volumes including heavy-duty vehicles occurs (Battarman, 2015; Battarman et al., 2015) may also play a role in these lower wintertime BC observations. Wang et al. (2017) recently showed that the proportion of non-volatile particle fraction (which may be expected to be predominantly comprised of BC) increased with ambient temperature. This factor also had a distinct weekday-weekend variation. The second BC-rich factor contained the highest concentrations of BrC1, BrC2 (see Table 1 for definition) and also contained organic ions that may be related to PAH naphthalene (one of the most abundant m/z 128). This factor appears to be a distinct mixing state of BC in which brown carbon was abundant. Healy et al. (2017) recently found a similar component that was attributed to biomass burning using the aethalometer data from the same location. On average, this BC-BrC factor contributed 17% of the total PM$_{2.5}$ mass at this location and Fig. 3d indicates that this factor was most abundant in the fall and winter while it was least abundant in the summer. Thus, recreational wood burning in the surrounding neighbourhoods is more likely the source rather than seasonal summertime wildfires that sometimes affect Ontario (Sofowote and Dempsey, 2015).

The third factor was high in crustal elements like Fe, Si and Ca. Other enriched elements included K, Mn, Zn, Pb and Ti as well as the chloride ion. This pattern is consistent with the re-suspension of road dust (Thorpe and Harrison, 2008). Pb and Zn may also be associated with tire and brake wear (Uijima et al., 2007). This RD factor contributed 6% to the total PM$_{2.5}$ mass on average. Wintertime abundance was high while concentrations were low in the warmer months (Fig. 3f) indicating that it was also associated with use of road salt and snow-clearing activities in winter.

The fourth factor was rich in the NH$_4^+$ and NO$_3^-$ ions as well as Cl$^-$ ion. Abundant organic ions in this factor include m/z 15, 18, 29, 31 and 44 indicating that organic aerosols may also be present in this factor. This particulate nitrate (pNO3) factor shows strong wintertime seasonality (Fig. 3h). The presence of Cl$^-$ also suggests that this factor is related to road salt which is used in winter. On average, it contributes 19% of the total PM$_{2.5}$ mass.

The fifth factor was relatively rich in m/z 29, 43, 45, 60 and 73 (Fig. 3i). While m/z 60 and 73 are also fragment ions associated with biomass burning markers such as levoglucosan, the presence of the m/z 45 strongly indicates that this factor was related to mono- and low molecular mass di-acids. This factor appears to have been extensively oxidized in the atmosphere given its summertime abundance (Fig. 3j), and concentrations in the cooler months were negligible, further suggesting that these ions could not be from levoglucosan which decomposes at a rapid rate in the summer (Hennigan et al., 2010; Mochida et al., 2010; Zhang et al., 2008). The sixth factor (OOA) also appears to be oxidized with a strong presence of oxidation marker ions m/z 18, 29, 44 and some 43 (Fig. 3k). The temporal pattern in Fig. 3l tracked the more acidic aerosols (fifth factor; Acid-A) with high summertime and low wintertime concentrations. This result suggests the importance of photochemistry in the generation of these oxidized aerosols. These two factors had an almost equal contribution for a total of 25% on average to PM$_{2.5}$ during the study period. Fig. 53 indicates that their peak summertime abundances appeared to mutually alternate from the first summer period in 2015 to the next in 2016.

The seventh and final mass-contributing factor was enriched in NH$_4^+$, SO$_4^{2-}$, V and some Se. Oxidation markers m/z 18 and 44 were also present (Fig. 3m). These species suggest that this factor could be attributed to particulate secondary sulphate (pSO4) derived from emissions of coal combustion. There was a distinct springtime abundance (Fig. 3n) that suggests that this factor may not be strictly related to energy generation with demand that usually peaks in summer especially considering that Ontario no longer uses coal for power generation (Ontario Ministry of the Environment and Climate Change, 2017). It should be noted that this factor is not expected to be dominated by primary sulfate emissions related to traffic since only ultralow S diesel fuel is used by heavy duty vehicles in Ontario. Thus, this factor may be related to local/regional metallurgical activities involving coking/coal combustion. The presence of V suggests No. 6 oil so shipping activity from Lake Ontario may also be contributing to this factor. Average total PM$_{2.5}$ mass contributions were relatively small at 10%. Finally, the eighth factor was rich in BC, aliphatic and aromatic hydrocarbon ions, i.e., m/z 27, 29, 41, 43, 53, 55, 57, 77, 79 and 91. This BC-HOA factor did not contribute appreciably to PM$_{2.5}$ mass since the sum of species (SHARP) PM is likely an underestimate of total ambient aerosol given that the SHARP is equipped with a heater and RH sensor along the inlet line which are activated to remove moisture when ambient RH > 35%. Volatile aerosols are also likely removed by this RH-controlled heat-drying.

3.3. Variation of the BC factor with PM$_{2.5}$ imbalance

The temporal trends of the ME2 factors were searched for the days wherein PM$_{2.5}$ imbalances were observed. The factor or group of factors that varied significantly were identified as being responsible for the imbalance. If a factor varied to similar extents when [PM$_{2.5}$]U.G. > [PM$_{2.5}$]H401, as when [PM$_{2.5}$]U.G. > [PM$_{2.5}$]H401, it could not be the main cause of the imbalance since there was an equal opportunity to be elevated when any imbalance event was observed. For an imbalance to occur, a single factor could ideally, have higher concentrations compared to other factors at both stations, or a group of factors could have small contributions that are each significantly different from zero. Also, for the [PM$_{2.5}$]U.G. > [PM$_{2.5}$]H401 case, it is possible that factors identified as causing that imbalance are likely at concentrations lower than their true/less-dispersed concentrations at the U.G. because all factors were determined with data from the H401 station where the effect of atmospheric dispersion is expected to be greater compared to distances closer to the source. Table 2 shows the average mass contributions of ME2-derived factors for the two cases of PM$_{2.5}$ imbalance between the H401 and U.G. stations and during periods of balance in percentages while Fig. 4 shows the ME2 factor contributions during the two PM$_{2.5}$ imbalance events. The first case in Fig. 4 represents imbalances when [PM$_{2.5}$]H401 > [PM$_{2.5}$]U.G., vice versa for the second case.

In general, the imbalances resulted from relatively higher average contributions of black carbon (BC) for the cases of [PM$_{2.5}$]H401 > [PM$_{2.5}$]U.G. BC-BrC and oxidized aerosol factors for the [PM$_{2.5}$]U.G. > [PM$_{2.5}$]H401 cases while secondary inorganic aerosol species (pSO4 and pNO3) were dominant during periods of PM$_{2.5}$ imbalance.

Table 2

<table>
<thead>
<tr>
<th>Factors</th>
<th>Case</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H401 &gt; U.G.</td>
</tr>
<tr>
<td>BC</td>
<td>40.2</td>
</tr>
<tr>
<td>BC-BrC</td>
<td>14.6</td>
</tr>
<tr>
<td>RD</td>
<td>7.7</td>
</tr>
<tr>
<td>pNO3 + pSO4</td>
<td>15.8</td>
</tr>
<tr>
<td>OOA + Acid-A</td>
<td>21.8</td>
</tr>
</tbody>
</table>
equivalece at both stations suggesting that these two factors are truly regional/long-range factors (see Table 2). These percentages give only a simplistic picture of the factors tied to imbalance since absolute factor contributions may not be significantly different. Other statistical tests must be used to compare the populations of data for the two imbalance events. To this end, non-parametric comparisons were done.

The Mann-Whitney-Wilcoxon non-parametric U test with z-statistics (Table S3) indicated that the BC-BrC and the Acid-A factors were not statistically different between the two imbalance cases while all the other factors were significantly different. The BC factor had the largest significant difference between the two imbalance cases \((Z_{\text{calc}} = 23.6 > z_{\text{crit}} = 1.96)\) and it generally drives the \([\text{PM}_{2.5}]_{\text{H401}} > [\text{PM}_{2.5}]_{\text{U.G.}}\) imbalance (Table 2, Fig. 4 and Fig. S5).

Also, Kruskal-Wallis ANOVA followed by Dunn multiple comparison test (Table S4) showed that for the \([\text{PM}_{2.5}]_{\text{H401}} > [\text{PM}_{2.5}]_{\text{U.G.}}\) imbalance case, only the BC factor was always significantly different from all other factors while for \([\text{PM}_{2.5}]_{\text{U.G.}} > [\text{PM}_{2.5}]_{\text{H401}}\), the OOA factor was always significantly different from all other factors. Furthermore, it is clear from Fig. 4 that except for the BC and BC-BrC factors, median values of most factors were not elevated (median around \(\leq 0.5 \mu g/m^3\)) when imbalances occur.

### 3.4. Parameterization of black carbon

Probing the diurnal variation of the BC factor (Fig. 5a; all other factor diurnal trends can be found in Fig. S4) reveals a distinct trend that peaks around 6:00 a.m. Hourly traffic count data averaged across the 12 lanes closest to the Wavetronix sensor probe are also included in Fig. 5b-d. Traffic categories \(C_2 \sim C_4\) were based on vehicle length. Thus, \(C_2\) are vehicles between 1 and 7.6 m in length, \(C_3\) are 7.6–15 m long vehicles and \(C_4\) are 15–36.5 m long vehicles. The length of the average sedan is 4.6 m, passenger utility truck is 6.2 m; maximum lengths of two-axle trucks and buses are 12.5 m, articulated bus is 25 m, tractor with regular semi-trailer is 23 m and tractor with tandem trailer is 25 m (Ontario, 2016). Fig. 5 shows that there is a somewhat apparent but not necessarily linear relationship between traffic category counts and the BC factor, indicating that other parameters may be critical to the variation.

The other two panels show the hourly average vehicular speed across all 12 lanes covered by the Wavetronix sensor (Fig. 5e) and hourly variations in the mixed depth of the planetary boundary layer (PBL) for the highway location estimated from HYSPLIT trajectories (Draxler and Rolph, 2010) for the period coinciding with the traffic count data availability.

The morning rush hour traffic count peaks roughly at 6:00 a.m. on weekdays for most vehicle types and this coincides with the BC factor contribution peak as well. After about 7:00 a.m. on weekdays, the BC factor begins to gradually diminish (Fig. 5a). A similar but much lower concentration pattern is observed in the weekend diurnal variations of the BC factor as well. Average vehicular speed also begins to drop after 6:00 a.m. on weekdays indicating the increasing traffic congestion during rush hour which is absent for the same time on weekends. The PBL depth shows lower heights in the mornings including the rush hour periods and gradually peaks just after mid-day followed by a gradual descent. This lowering of the mixed layer height serves to trap and exacerbate the pollutant concentrations. When \(C_2\) and \(C_4\) counts peak before noon, the BC factor concentrations have already declined from their morning rush hour maxima likely due to the increase in mixing depth, typically higher wind speeds and relatively higher traffic speeds that dilute out the concentrations. All of these additional parameters affect the relationship between the BC factor (> 90% BC6 by mass) and traffic category counts. The objective was then to parameterize this relationship as best as possible.

Thus, the data on imbalance events were used to identify the time periods when the highway roadside station was expected to be impacted primarily by highway traffic, i.e., when \([\text{PM}_{2.5}]_{\text{H401}} > [\text{PM}_{2.5}]_{\text{U.G.}}\) as defined in equation (5). Two regression model types were assessed wherein the hourly traffic category count data \((C_2 \sim C_4)\) and their average vehicular speed (AVS) across the 12 lanes of coverage, mixed depth (MD), cosine of the trigonometrically-converted wind direction \((\theta = \lambda \cdot WD,\) where \(\lambda = 90^\circ\) if WD \(\leq 90^\circ\), otherwise \(\lambda = 450^\circ\), so that all negative values indicate west and all positive values are east), wind speed (WS), and ambient temperature (ATEM) were used as explanatory variables for the corresponding BC6 (BC at 880 nm) data used as the dependent variable. A multiple linear regression was performed, but poor \(R^2\) results suggested that a linear relationship was inappropriate (see example in Table S5). Thus, non-linear regressions were performed with two methods. These two methods differed in the way they performed their gradient descent (Friedman and Popescu, 2004).

The first method used was the generalized reduced gradient model (GRG; Brown, 2001; Smith and Lasdon, 1992). The GRG model (Frontline Solvers, 2017) is particularly suited to convex nonlinear optimization problems for which true global solutions can be determined. GRG allows for flexibility in the formulation of the relationship between dependent and explanatory variables and iteratively minimizes the sum of squares between the observed and model-fitted values. Placement of explanatory variables (as numerators or denominators) in these relationships was guided by a combination of presumptions and successive trials that explained the most variance. Many such formulations were tested and Table 3 below presents summary outputs of four different nonlinear formulations of the regression between BC6 and the...
aforementioned parameters for imbalance events \([\text{PM}_{2.5}]_{\text{H401}} > [\text{PM}_{2.5}]_{\text{U.G.}}\) largely to show the effect of stepwise additions of new parameters on the proportion of variance explained \((R^2)\). In all cases, automatic scaling and a precision of 0.000001 were used while the convergence criterion was set at \(1 \times 10^{-9}\). Finally, most of the solved-for coefficients were unconstrained for maximum \(R^2\). Only the exponents were sometimes constrained between 0.5 and 1 so the solutions could have a realistic interpretation. GRG was run in Microsoft Excel (Office 2010; Solver add-in) until it indicated it had converged to or found a solution. A true global solution may however not be feasible in this analysis because the dependent variable \([\text{BC6}]\) is not fully convex (or more correctly in this case, concave), so it is likely that these solutions represent local minima. Summary reports of the solutions can be found in Tables S6 – S9.

The second method belongs to a family of large data mining algorithms called machine learning wherein the input data is subdivided...
into a training/learning and test set. It involved the use of ensembles of rules in a multi-member regression wherein a penalty (mainly, the lasso) is introduced for regularization or model stabilization (Friedman, 2012; Friedman and Popescu, 2008). More details on the kinds of regularization used in rule-fitting can be found elsewhere (Friedman et al., 2010). Rules are generated from the data using decision trees estimated from large decision trees. The importance of a rule is dependent on its validity/truth of the rules. Base learners can range from simple functions of the independent variables to multi-node rules from large decision trees. The importance of a rule is dependent on its scale which is in turn dependent on the support of the rule which is reported in the model output and is optimally removed from 0 to 1.

The algorithm (RuleFit; Friedman, 2012) was run in the statistical program R (http://cran.usstat.utoronto.ca/, 2017) to generate and select the fitted rules as well as develop the initial model in the training set with a subsequent validation stage. An intrinsic benefit of this algorithm is the extensive data cross-validation wherein measures of model accuracy (average absolute error; AAE) are re-computed by applying model outputs derived from a multiple (K-fold; Friedman and Popescu, 2004) partitioning of the training set on the test set. Rule based ensemble regressions have been shown to yield more accurate results (lower AAE and error rate) when compared with other data mining (Friedman, 2012), and regression methods (Haghighi et al., 2016). In this work, the model was allowed to fit both linear functions of the independent variables and rules generated as discussed above, the model selected was for regression, the specified maximum number of rules for regression fitting was 2000 while the specified maximum number of terms in the final model was 500, and a tree size of 4 was used so as to allow for reasonable interaction among variables. Cross validation was done with a fold value of K = 200 and main model criteria output were the AAE, cross-validated AAE and root-mean-square error (RMSE), the rules selected, their coefficients and supports. Other outputs obtained were the relative importance of variables, variable interaction summary plots, single and dual variable partial dependencies of the response.

With GRG, the formulation of the relationship between BC6 and the traffic-related and meteorological parameters in the fourth column of Table 3 gave the best explained variance, i.e., R² value (highlighted in boldface). This relationship only concerns the case of the PM2.5 imbalance where [PM2.5]H401 > [PM2.5]U.G.. The relationship of average contributions of the three categories of vehicles can be estimated from their coefficients and it indicates that overall, the C2 and C4, i.e., heavy-duty diesel-powered vehicles dominate the relationship with BC6 while C2 is negligible. The negative C2 coefficient is likely due to periods such as weekend mornings where C2 is rapidly increasing but values of the BC factor are declining (see Fig. 5a and b).

The rule-ensemble nonlinear method (Rulefit results: AAE = 530.9 ± 19.08; Cross-validated AAE = 540.8; Cross-Validated RMSE = 679.8; number of terms = 135) also indicates that C2 and C4 are the most important traffic-related variables (Fig. 6a) and the rules involving these two variables tend to be the most important ones (Table 4). Variable importance can also be gleaned from the partial dependence of the response variable on individual regressors. Fig. S6 shows the partial dependence of BC6 on the relevant ranges of the independent variables. Apart from revealing largely nonlinear relationships of BC6 with these variables, it can be seen that the range of the y-axis decreases in the order of Cos θ > C3 > WS = MD = C4 > AVS > ATEM = C3, indicating that the last two variables have negligible impact on BC6 during these imbalance events. Fig. 6b and Fig. S7 are examples of joint partial dependence of BC6 on C2 and mixed depth (MD).

<table>
<thead>
<tr>
<th>BC6</th>
<th>bC2, C3, AVS, MD, ATEM, WS, Cos θ</th>
<th>( R^2 )</th>
<th>df</th>
<th>Critical t</th>
<th>PI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bC2</td>
<td>bC3</td>
<td>bAVS</td>
<td>bWS</td>
<td>bATEM</td>
</tr>
<tr>
<td></td>
<td>5.20</td>
<td>−0.20</td>
<td>0.52</td>
<td>0.70</td>
<td>0.000083</td>
</tr>
<tr>
<td></td>
<td>12.37</td>
<td>12.00</td>
<td>15.16</td>
<td>14.62</td>
<td>0.000083</td>
</tr>
<tr>
<td></td>
<td>83.54</td>
<td>8.88</td>
<td>78.43</td>
<td>4.98</td>
<td>0.000083</td>
</tr>
</tbody>
</table>

Table 3 Nonlinear regression modelling results for the relationship of BC6 with traffic count-related (C2-C4, AVS) and meteorological parameters (MD, ATEM, WS, Cos θ) for the imbalance case when [PM2.5]H401 > [PM2.5]U.G.. The rightmost parameters gave the best \( R^2 \) value.

Fig. 6. Plots of rule-ensemble nonlinear modelling (RuleFit) results showing: (a) relative variable importance with respect to BC6 during imbalance events; (b) joint partial variable dependency of BC6 on C2 and mixed depth (MD).
respectively clearly indicating that BC6 is high when C3 is high and mixed depths are low and winds are coming from the north-west. Also, of all the traffic-related variables, C3 interacts the most with other variables (Figs. S8 and S9).

Other researchers have pointed to potential significant contributions from light-duty gasoline vehicles to BC concentrations with a significant < 75 nm particle mode (Liggio et al., 2012), likely from the increasing importance of gasoline direct injection (GDI) powered vehicles that may generate more BC particles (Liggio et al., 2012). However, the same kinds of contributions from light duty vehicles were not observed in this study, possibly because compared to our current data coverage was poor; i.e., the period for which count data is available. For the work presented herein, the BC parameterization does not account for the contributions of precipitation that may result in a wash-out of BC.

Fig. 7 shows the relationships between the original and modelled BC6 concentrations using the coefficients of parameters from the fourth column of Table 3 for the relevant PM2.5 imbalance events, i.e., when both PM2.5 and BC are likely to be greater at the highway station. Fig. 8 shows similar plots but with rules and their coefficients partially listed in Table 4. There is good agreement (slope and R²) between the original and estimated BC6 values.

There are several possibilities why the models do not totally explain the observed BC6 concentrations during imbalance. First, there are other urban/regional sources of BC impacting the area inclusive of the highway that result in background concentrations that in turn cannot be accounted for by highway traffic alone. Table 2 shows that the BC factor makes up 20% of the average PM mass during periods of PM imbalance between the two stations but doubles to 40% when PM2.5 > PM2.5U.G. and Fig. S5 shows that during these imbalance periods, there is a > 60% increase in the average contributions of the BC factor over equivalence periods (170% increase in the median), so the highway traffic can only directly explain up to 63% of all the BC at the roadside station. This may help explain why the slopes of Figs. 7a and 8a are ~ 0.66–0.74 (though, the cross-validated RuleFit slope 0.64 in Fig. 8a appears most realistic), and also, the presence of a y-axis intercept for the regression line between the original and estimated BC6.

Jeong et al. (2013) reported long-range advected BC that could potentially impact rural locations in southern Ontario and would produce background concentrations in Toronto. More recently, Healy et al. (2017) found relatively low but seasonally-varying regionally-advected BC concentrations at Hanlan’s Point; a background station on Toronto Island with no local sources. Such background BC concentrations may help explain the events where the originally measured BC is greater than the estimated one.

Secondly, the BC parameterization does not account for the confounding effects of precipitation that may result in a wash-out of BC. Precipitation events will cause the observed BC concentrations to be lower than the modelled value.

Finally, this relationship could be re-optimized in the future when more traffic data are available (note the absence of most summertime results due to Wavefrontix operational failures). More data may lead to better agreements (i.e., better R² and lower intercept values) between the original and estimated BC concentrations. It would be possible to add these traffic count variables into the initial ME2 input data matrix when more traffic count data is available. For the work presented herein, the traffic data coverage was poor; i.e., the period for which validated traffic data (about 150 days i.e., ~ 3600 Hrs) was available and overlapped with the other data was only 2616 Hrs in total. This data acquisition rate represents ~ 40% coverage of the metals data and ~ 11% of the ACSM data.

4. Conclusions

The objective of this study was to determine the causes and proportion the sources of PM2.5 imbalance between a roadside station and another site relatively farther away from an extremely busy highway. The benefit of identifying the causes of this imbalance is that effective
and targeted pollution control strategies can then potentially be developed. In this case, the major cause of the imbalance appears to be heavy-duty vehicular traffic on the highway. The chemical signature of this particulate matter pollution as determined by high-time resolution factor analysis is a black carbon (BC) factor. By mass, this BC factor is almost entirely comprised of BC and is most elevated during the morning rush hour period at this location.

Parameterization of meteorological and traffic-related variables to explain this BC pollution indicates that heavy-duty vehicles over ≥ 7.6 m in length are largely responsible for this pollutant. Heavy-duty vehicles of this size are usually powered by diesel fuel. The seasonal variability of carbon-rich aromatic species in diesel fuel composition also mirrors the BC trend. Conservatively, a minimum 400,000 people traverse this portion of the highway daily and roughly one-seventh of this number during the morning rush hour alone. From a policy perspective, the quickest way to control exposure to local roadway BC is to design mitigation programs that target these two classes of vehicles (C3 and C4) since other emissions that target these two classes of vehicles (C3 and C4) since other emissions

Acknowledgements

The authors gratefully acknowledge the NOAA Air Resources Laboratory (ARL) for the provision of the HYSPLIT transport and dispersion model used in this application for the generation of mixed depths of the planetary boundary layer. We want to thank Dr. Pentti Paatero for useful discussions at the start of the data analysis work. The authors also acknowledge Dr. Philip Croteau of Aerodyne Research Inc. for the provision of the ACSM Error Calc algorithm module used in this work.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.atmosenv.2017.10.063.

References
