Science of the Total Environment 521–522 (2015) 19–26

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Forecasting of VOC emissions from traffic and industry using classification and regression multivariate methods

Andreja Stojić [a,*], Dimitrije Maletić [a], Svetlana Stanišić Stojić [b], Zoran Mijić [a], Andrej Šoštarić [c]

[a] Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia
[b] Singidunum University, Danijelova 32, 11010 Belgrade, Serbia
[c] Institute of Public Health Belgrade, Bulevar Despota Stefana 54, 11000 Belgrade, Serbia

HIGHLIGHTS

- Receptor models were applied for the purpose of VOC source apportionment.
- MVA methods were used for forecasting contributions from traffic and industry.
- Forecast was based on inorganic pollutant concentrations and meteorological data.
- Predicted values were consistent with the results of receptor modeling.
- The highest forecast accuracy was achieved with relative error of only 6%.

ARTICLE INFO

Article history:
Received 19 December 2014
Received in revised form 21 March 2015
Accepted 22 March 2015
Available online 28 March 2015

Editor: P. Kassomenos

Keywords:
VOC
PTR-MS
Receptor modeling
Forecasting
MVA

ABSTRACT

In this study, advanced multivariate methods were applied for VOC source apportionment and subsequent short-term forecast of industrial- and vehicle exhaust-related contributions in Belgrade urban area (Serbia). The VOC concentrations were measured using PTR-MS, together with inorganic gaseous pollutants (NO_x, NO, NO_2, SO_2, and CO), PM_{10}, and meteorological parameters. US EPA Positive Matrix Factorization and Unmix receptor models were applied to the obtained dataset both resolving six source profiles. For the purpose of forecasting industrial- and vehicle exhaust-related source contributions, different multivariate methods were employed in two separate cases, relying on meteorological data, and on meteorological data and concentrations of inorganic gaseous pollutants, respectively. The results indicate that Boosted Decision Trees and Multi-Layer Perceptrons were the best performing methods. According to the results, forecasting accuracy was high (lowest relative error of only 6%), in particular when the forecast was based on both meteorological parameters and concentrations of inorganic gaseous pollutants.

1. Introduction

Volatile organic compounds (VOC) comprise a diverse group of species which are of concern due to their potentially detrimental impact on human health and the environment. Under sufficiently conducive meteorological conditions, they are important precursors in the formation of ozone, the abundant and reactive gaseous pollutant, capable of inducing oxidative damage to living cells (Kampa and Castanas, 2008). In addition, several VOC species such as styrene and benzene have been identified as toxic or mutagenic, while epidemiological evidence indicates that repetitive daily or intermittent exposure is associated with numerous adverse health effects, mainly respiratory and hearth disorders (Musselman and Korfmacher, 2014; Hsieh and Tsai, 2003). As regards environmental issues, the significant impact of VOC on climate change is observed in spite of their low concentrations in ambient air, and arises from their ability to form secondary aerosol and their properties as greenhouse gases (Chin and Batterman, 2012).

The ubiquity of VOC results from both biogenic and anthropogenic emissions, whereas the latter often dominate in heavily populated areas and are associated with vehicle transport, industrial activities, fossil fuel refining and distribution, biomass burning, solvent usage, *etc.* (Lee et al., 2002; Na et al., 2004). The abundance and spatial distribution of gaseous pollutants originating from remote emission sources mostly depend on their atmospheric lifetimes (Jobson et al., 1999), whereas, in the case of locally generated pollution, this relationship is no longer sustained, and VOC levels and variability are mainly controlled by emission rates and meteorological factors (Liu et al., 2012).

---

* Corresponding author at: Pregrevica 118, 11080 Belgrade, Serbia.
E-mail addresses: andreja.stojic@ipb.ac.rs (A. Stojić), dimitrije.maletic@ipb.ac.rs (D. Maletić), sstanisic@singidunum.ac.rs (S. Stanišić Stojić), zoran.mijic@ipb.ac.rs (Z. Mijić), andrej.sostaric@zdravlje.org.rs (A. Šoštarić).

http://dx.doi.org/10.1016/j.scitotenv.2015.03.098
0048-9697/

Apart from being an important regional traffic hub with a population of 1.6 million residents, the capital of Serbia, Belgrade, and its suburban area is home to network of coal-fired power plants and different industrial facilities, such as the petrochemical complex, chemical plant, and oil refinery. In such complex urban environments with the prevalence of local emission sources, meteorological conditions play a significant role in the VOC mixing and distribution. The methods which simulate the variations of VOC emissions with sufficient reliability, based solely on meteorological data, can be used to forecast temporal distribution of VOC species, which is essential for development of efficient abatement strategies (Liu et al., 2012).

In this study several multivariate (MVA) methods were employed to assess the impact of traffic- and industry-related sources on VOC levels in Belgrade urban area, and predict their contribution dynamics. The petrochemical/chemical industry (PC) and vehicle exhaust emissions (VE) present one of the most significant emission sources. Their contributions were estimated using widely applied receptor models, Positive Matrix Factorization (PMF) and Unmix, based on the assumption that, in a complex VOC mixture, species emitted from the same source are statistically interrelated (Song et al., 2008). Subsequently, classification and regression MVA methods were applied in order to predict the source contribution dynamics on the basis of meteorological dataset and concentrations of inorganic gaseous pollutants (IG) — NO_x, NO, NO_2, SO_2 and CO. The supervised learning algorithms for classification and regression analyses were specifically designed within Toolkit for Multivariate Analysis (TMVA) (Hoecker et al., 2007) within the ROOT framework (Brun and Rademakers, 1997), for extensive data processing in high-energy physics, but their applications are not restricted to these requirements (Maletić et al., 2014). The best performing MVA methods were Boosted Decision Trees (BDT and BDTG), designed for the purpose of MiniBooNE neutrino experiment (Yang et al., 2005), and based on complex method of cuts, and Multi-Layer Perceptrons (MLP), based on artificial neural networks (ANN) (Rojas, 1996).

Nowadays, the forecasting of air pollutant concentrations is an essential issue in environmental research due to a wide range of potential benefits. Besides providing information for early public warnings to the susceptible populations, as well as assistance in the assessment of regulation policies, the accurate and reliable forecast could be useful for development of preventive approaches and considerable reduction in the number of measurement sites over the area. This study reveals that presented MVA methods can be successfully used for forecasting the contributions of different emission sources in the investigated area.

2. Materials and methods

The measurement site is located at the Institute of Public Health in Belgrade (44°49′ N, 20°28′ E), in the urban canyon street with heavy and slow traffic. Proton Transfer Reaction Mass Spectrometer (Standard PTR-quad-MS, Ionicon Analytik, GmbH, Austria) was used for on-line measurements of concentrations of 36 VOC-related masses in the period from January 22nd to March 24th 2014. A detailed description of the method is given elsewhere (Lindinger et al., 1998; de Gouw and Warneke, 2007). The inlet of the instrument, 3 m heated (70 °C) silcosteel line inner diameter 3 mm, was placed 3 m above ground. VOC data, with 0.5 s dwell time, and five control parameters (m/z 21, m/z 25, m/z 30, m/z 32, and m/z 37) were obtained in 24 s cycles.

Drift tube parameters included: pressure, ranging from 2.08 to 2.11 mbar; temperature, 60 °C; voltage, 600 V; E/N parameter, 145 Td providing reaction time of 90 μs. The count rate of $H_3O^+H_2O$ was 1 to 7% of the $5.1 \cdot 10^6$ counts s^{-1} count rate of primary H_3O^+ ions. The calibration was done according to Taipale et al. (2008). For this purpose, TO-15 Supelco gas mixture (m/z 57, m/z 79, m/z 93, m/z 107, and m/z 121) was diluted with ASGU 370-p HORIBA system zero air to five concentrations in the range from 0.5 to 100 ppb. Normalized sensitivities were in the range from 6.2 to 14.3 npcs ppb^{-1}. Detection limit of 1-h averaged VOC concentrations was less than 0.5 ppb, except for methanol (2.0 ppb) and acetone (1.1 ppb). Key m/z-signals in the dataset, identified using a method developed by Galbally et al. (2008), were observed for 29 out of 36 masses, and used for further analysis.

The concentrations of IG, PM_{10}, and meteorological data (atmospheric pressure, temperature, humidity, precipitation, wind speed and direction) were obtained from the automatic monitoring station at the measurement site (Institute of Public Health Belgrade).

US EPA Unmix 6.0 (USEPA, 2007) and Positive Matrix Factorization (Version 3.0) (USEPA, 2008) receptor models were applied to the 1169 observations of 1-hour-averaged concentrations of 29 species in order to identify emission sources (Table 1). The usages of these models as well as the theoretical background are detailed in literature (Henry, 2003). Briefly, Unmix is based on an eigenvalue analysis and does not allow down weighting of individual data points (Henry, 1997), while PMF decomposes a matrix of ambient data into two matrices representing source contribution and source profile (Paatero and Tapper, 1994). A value equal to the half of the method detection limit (DL) for each variable was used for concentrations below the DL. The number of pollutants selected as Unmix and PMF input variables was chosen using a combination of trial and error with the general goal of maximizing the number of input variables that produced feasible and physically interpretable solutions and following additional fit diagnostics criteria (Chan et al., 2011). After selecting the base run, 100 bootstrap runs with R^2-value of 0.6 were performed to evaluate the uncertainty of the PMF resolved profiles. In addition, PMF was run with different Fpeak values to explore the rotational freedom and reported results were for its value adjusted to 0.2.

Statistical analyses, including bivariate polar plot and bivariate cluster (k-means clustering, grouping similar conditions together) analysis, were performed with the statistical software environment R (Team, 2012), using the Openair package (Carslaw and Ropkins, 2012). The

Table 1
Basic statistics for measured parameters: VOC related masses [ppb], NO_x, NO_2, NO, SO_2 [μg m^{-3}], and CO [mg m^{-3}] concentrations.

Parameter	Mean	Median	Min	Max	10th	90th	St. dev.
NO_x	149.69	121.01	11.24	912.42	37.14	301.52	116.19
NO_2	62.94	54.96	8.86	239.09	21.36	115.61	38.17
NO	86.76	58.67	1.19	673.33	11.12	205.61	88.32
CO	0.68	0.59	0.16	3.42	0.36	1.10	0.38
SO_2	21.54	17.43	3.85	236.46	7.74	39.94	16.70
Ethylbenzene	2.31	1.76	<DL	36.93	0.37	4.63	2.41
mp-Xylene	8.99	6.86	<DL	124.62	1.41	17.25	9.52
o-Xylene	1.88	1.46	<DL	18.52	0.28	3.79	1.72
m/z 41 (propylene)	1.89	1.66	0.34	19.88	0.80	2.94	1.43
m/z 43	4.87	4.42	1.27	28.40	2.32	7.84	2.52
m/z 45 (acetaldehyde)	5.04	4.29	1.31	38.85	2.28	7.92	3.53
m/z 47 (ethanol)	7.81	3.02	<DL	177.33	<DL	17.39	15.55
m/z 57 (MTBE)	1.94	1.63	0.26	28.82	0.75	2.92	1.93
m/z 59 (acetone)	7.16	5.67	<DL	30.79	1.61	15.57	5.42
m/z 61 (acetic acid)	4.85	4.48	1.36	25.98	2.58	7.24	2.36
m/z 71	0.63	0.58	0.11	5.32	0.29	0.94	0.39
m/z 73	0.87	0.78	0.17	7.89	0.46	1.32	0.52
m/z 75	1.28	0.97	0.26	45.29	0.54	1.80	2.09
m/z 79 (benzene)	1.35	1.10	0.05	6.95	0.50	2.44	0.90
m/z 81	0.81	0.68	0.07	9.88	0.25	1.33	0.78
m/z 83	14.4	10.16	0.37	92.44	3.05	31.65	11.99
m/z 85	9.43	6.73	0.17	65.12	1.96	20.52	7.85
m/z 87	2.53	2.20	0.49	12.36	1.03	4.55	1.48
m/z 93 (toluene)	3.36	2.63	0.19	29.66	1.15	6.11	2.62
m/z 99	0.49	0.43	<DL	2.32	0.17	0.91	0.31
m/z 101	1.25	1.05	0.37	5.34	0.58	2.33	0.71
m/z 105 (styrene)	0.53	0.43	0.11	16.39	0.22	0.76	0.81
m/z 121 (C9 aromatics)	4.32	3.13	0.28	24.96	1.19	8.59	3.88
m/z 137 (monoterpenes)	0.89	0.65	<DL	8.50	0.19	1.67	0.90

DL — detection limit.