Supplementary data for article:

Sostaric, A.; Stojic, S. S.; Vukovic, G.; Mijic, Z.; Stojić, A.; Gržetić, I. Rainwater Capacities for BTEX Scavenging from Ambient Air. *Atmospheric Environment* **2017**, *168*, 46–54.

https://doi.org/10.1016/j.atmosenv.2017.08.045

Rainwater capacities for BTEX scavenging from ambient air

A. Šoštarić[1], S. Stanišić Stojić[2], G. Vuković[3], Z. Mijić[3], A. Stojić[3] and I. Gržetić[4]

1 Institute of Public Health Belgrade, Bulevar Despota Stefana 54, 11000 Belgrade, Serbia
2 Faculty of Physical Chemistry, University of Belgrade, Studentski trg 12-16, 11000 Belgrade, Serbia
3 Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia
4 Faculty of Chemistry, University of Belgrade, Studentski trg 12-16, 11000 Belgrade, Serbia

Atmospheric Environment

Supplementary material

Physico-chemical methods for rainwater analysis

The analyses of rainwater physico-chemical properties included determination of the major inorganic anions (F^-, Cl^-, SO_4^{2-}, NO_2^- and NO_3^-), dissolved cations (Na^+, NH_4^+, K^+, Ca^{2+} and Mg^{2+}), total organic carbon, electrical conductivity, UV extinction, turbidity and pH, in accordance with the standard methods (US EPA 300.1:1993, EN ISO 14911:1998, ISO 8245:1999, EN 27888:1993, SMEWW 19th method 5910 B, US EPA 180.1:1993, EN ISO 10523:2008, respectively).

Sample preparation referring to sample aliquot filtration through a 45 µm cellulose membrane syringe filter (Econofilter, Agilent) was conducted only for the determination of major inorganic anions and dissolved cations.

Inorganic anion content (F^-, Cl^-, SO_4^{2-}, NO_2^- and NO_3^-) was determined by the use of EPA method 300.1:1997 (Determination of inorganic anions in drinking water by ion chromatography – USEPA, 1997). Chromatographic separation and quantification of separated species was performed by the use of Methrom 761 Compact IC equipped with precolumn, Metrosep A Supp 5 - 250/4.0 separation column with suppressor, conductivity detector and mixture of sodium carbonate and sodium bicarbonate as mobile phase.

Dissolved cation content (Na^+, NH_4^+, K^+, Ca^{2+} and Mg^{2+}) was determined according to EN ISO International Standard 14911:1999 (Water quality - Determination of dissolved Li^+, Na^+, NH_4^+, K^+, Mn^{2+}, Ca^{2+}, Mg^{2+}, Sr^{2+} and Ba^{2+} using ion chromatography - Method for water and waste water - ISO, 1999). Chromatographic separation and quantification of separated species was performed by the use Methrom 761 Compact IC, using Metrosep C 4 - 100/4.0 separation column and mixture of dipicolinic and nitric acid as mobile phase.

Quality assurance/quality control measures for the samples analyzed by ion chromatography included duplicate samples, laboratory blanks, fortified laboratory blanks and field blanks.

Total organic carbon content was determined according to EN ISO International Standard 8245:2007 (Water quality - Guidelines for the determination of total organic carbon (TOC) and dissolved organic carbon (DOC) - ISO 2007) by the use of PPM LABTOC Analyser, based on organic carbon oxidation to carbon dioxide by UV light and sodium persulphate, quantified by the use of solid state infrared detector.

Electrical conductivity was determined according to EN International Standard 27888:2009 (Water quality - Determination of electrical conductivity - EN 2009) and Metrohm 712 conductometer.

UV extinction was determined according to Standard Methods for the Examination of Water and Wastewater 5910 B - UV-Absorbing Organic Constituents (SMEWW, 2000) and Analytik Jena, SPECORD

50, UV VIS spectrophotometer.

UV extinction was determined according to Standard Methods for the Examination of Water and Wastewater 5910 B - UV-Absorbing Organic Constituents (SMEWW, 2000) and Analytik Jena, SPECORD 50, UV VIS spectrophotometer.

Turbidity was determined according to EPA method 180.1:1993 (Determination of Turbidity by Nephelometry - USEPA, 1993) and HACH 2100N Laboratory Turbidimeter with formazin polymer as reference suspension.

The **pH** was determined according to EN ISO International Standard 10523:2016 (Water quality - Determination of pH - ISO 2016) and Metrohm 781 pH-meter.

All methods are accredited according to ISO 17025 and subjected to strict quality assurance, quality control programme and validation performed for rain samples.

Multivariate regression methods

Weka (Frank et al., 2005) is a collection of machine learning algorithms for data mining tasks that contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. All the machine learning techniques used in this paper are implemented in Weka so that they will be easily and fairly compared to each other. In addition to methods implemented in Weka 3.8 core, other regression methods of the packages found in Weka unofficial repository are used in this paper (Weka Sourceforge, 2016). Brief method descriptions are listed below.

Alternating model tree method grows an alternating model tree by minimizing squared error (Frank et al., 2015). It contains splitter and prediction nodes, using simple linear regression functions as opposed to constant predictors at the prediction nodes. Moreover, additive regression using forward stagewise modeling is applied to grow the tree rather than a boosting algorithm. The size of the tree is determined using cross-validation. The results show that the model achieves significantly lower squared error than standard model trees.

Conjunctive rule is the machine learning algorithm known as inductive learning. The goal of rule induction is generally to induce a set of rules from data that captures all generalizable knowledge within that data, and at the same time, to be as small as possible (Cohen, 1995).

Decision stump is a one-level decision tree where the split at the root level is based on a specific attribute/value pair (Zhao and Zhang, 2008).

Decision table builds a decision table majority classifier. It evaluates feature subsets using best-first search and can use cross-validation for evaluation (Witten and Frank, 2005).

Elastic net is a generalization of the lasso method for linear regression which uses regularization, i.e. penalization methods that introduce additional constraints into the optimization of a predictive algorithm that bias the model toward lower complexity (Zou et al., 2006).

Gaussian processes is a collection of random variables, any finite number of which have (consistent) joint Gaussian distributions. It uses lazy learning and a measure of the similarity between points to predict the value for a point from training data (Rasmussen, 2006).

IBk is a type of instance-based learning (*k* nearest neighbors – k-NN) where the function is only approximated locally and all computation is deferred until classification. The output value is the average of its *k* nearest neighbors (Aha et al., 1991).

IBkLG is an improvement of K-NN method by introducing weighted distance based on the negative logarithm or a Gaussiankernel (IBkLG, 2015).

Isotonic regression implements the method based on pair-adjacent violators approach. It is learner which picks the attribute that results in the lowest squared error (Witten and Frank, 2005).

K* is an instance-based classifier, that is, the class of a test instance is based upon the class of training instances similar to it, as determined by some similarity function (Cleary and Trigg, 1995). It contrasts to other instance-based learners, it uses an entropy-based distance function.

Least median squares regression is utilization of the existing Weka Linear Regression class to form predictions. The Least Squares regression functions are generated from random subsamples of the data. The least squares regression with the lowest median squared error is chosen as the final model (Rousseeuw and Leroy, 2005).

Linear regression method minimizes the sum of the squared difference between the observed and predicted values, and creates a line for optimal data separation (Shi and Tsai, 2002). Models were formed by random subsamples and the best performing model was selected according to Akaike information criterion (AIC).

Locally weighted learning (LWL) uses an instance-based algorithm to assign instance weights where a linear regression model is fit to the data based on a weighting function centered on the instance for which a prediction is to be generated (Frank et al., 2003). The resulting estimator is nonlinear because the weighting function changes with every instance to be processed.

M5P is based on decision trees, however, instead of having values at tree's nodes, it contains a multivariate linear regression model at each node (Graczyk et al., 2009). The input space is divided into cells using training data and their outcomes, then a regression model is built in each cell as a leaf of the tree.

M5 rules generates a decision list for regression problems using separate-and-conquer. It builds a model tree in each iteration using M5 and makes the "best" leaf into a rule (Wang and Witten, 1996).

Multilayer perceptron is a feed forward neural network model with one or more hidden layers between input and output layer (Haykin, 1994). It is trained by the back-propagation algorithm that uses gradient descent to minimize error and adjust the weights to each connection between the hidden and output layer.

Pace regression is provably optimal under regularity conditions when the number of coefficients tends to infinity. It consists of a group of estimators that are either overall optimal or optimal under certain conditions (Wang and Witten, 2002).

Random forest is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process (Breiman, 2001; Zhao and Zhang, 2008). The prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Random forest generally exhibits a substantial performance improvement over the single tree classifier such as CART and C4.5. It yields generalization error rate that compares favorably to Adaboost, yet is more robust to noise.

Random tree constructs a tree that considers K randomly chosen attributes at each node (Zhao and Zhang, 2008). In this context, “at random” means that each tree in the set of trees has an equal chance of being sampled. The method performs no pruning. It has an option to allow estimation of class probabilities based on a hold-out set (backfitting).

Radial base function (RBF) is an artificial neural network which implements a normalized Gaussian radial basis function network. It uses the k-means clustering algorithm to provide the basis functions and learns either a logistic regression (discrete class problems) or linear regression (numeric class problems) on top of that. Symmetric multivariate Gaussians are fit to the data from each cluster. It standardizes all numeric attributes to zero mean and unit variance (Frank, 2014).

REP tree is a fast decision tree learner. It builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting) (Kalmegn, 2015).

Simple linear regression picks the attribute that results in the lowest squared error.

SMOreg is a Support vector machine implemented using the Sequential Minimal Optimization Regression (SMOreg) algorithm (Shevade et al., 1999). By means of non-linear function, the input data are mapped into a high dimensional feature space, where linear regression is performed.

Figures

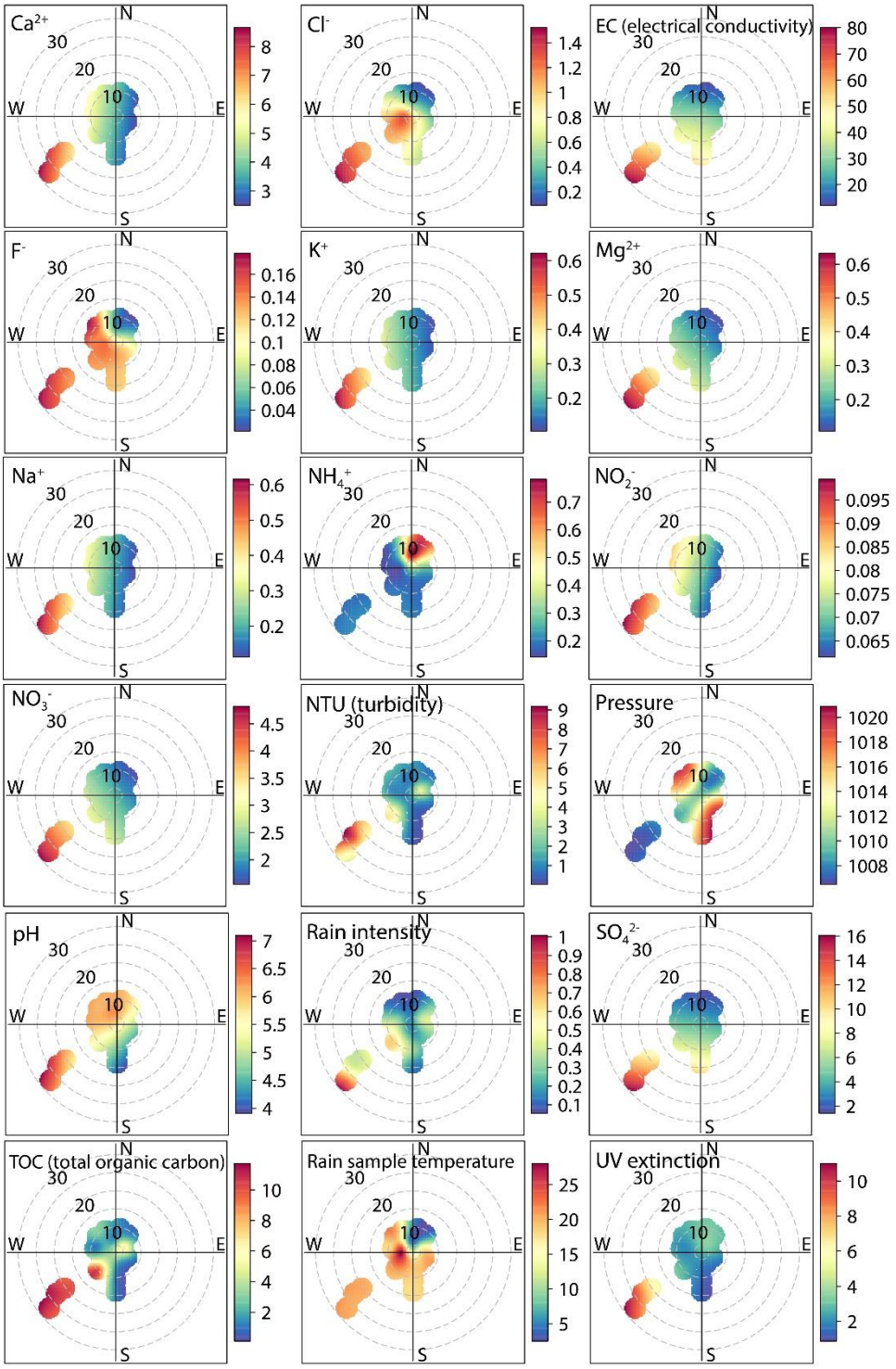

Figure S1. Measured concentrations/parameters dependence on wind speed [m s^{-1}] and direction (units are presented in Table S2).

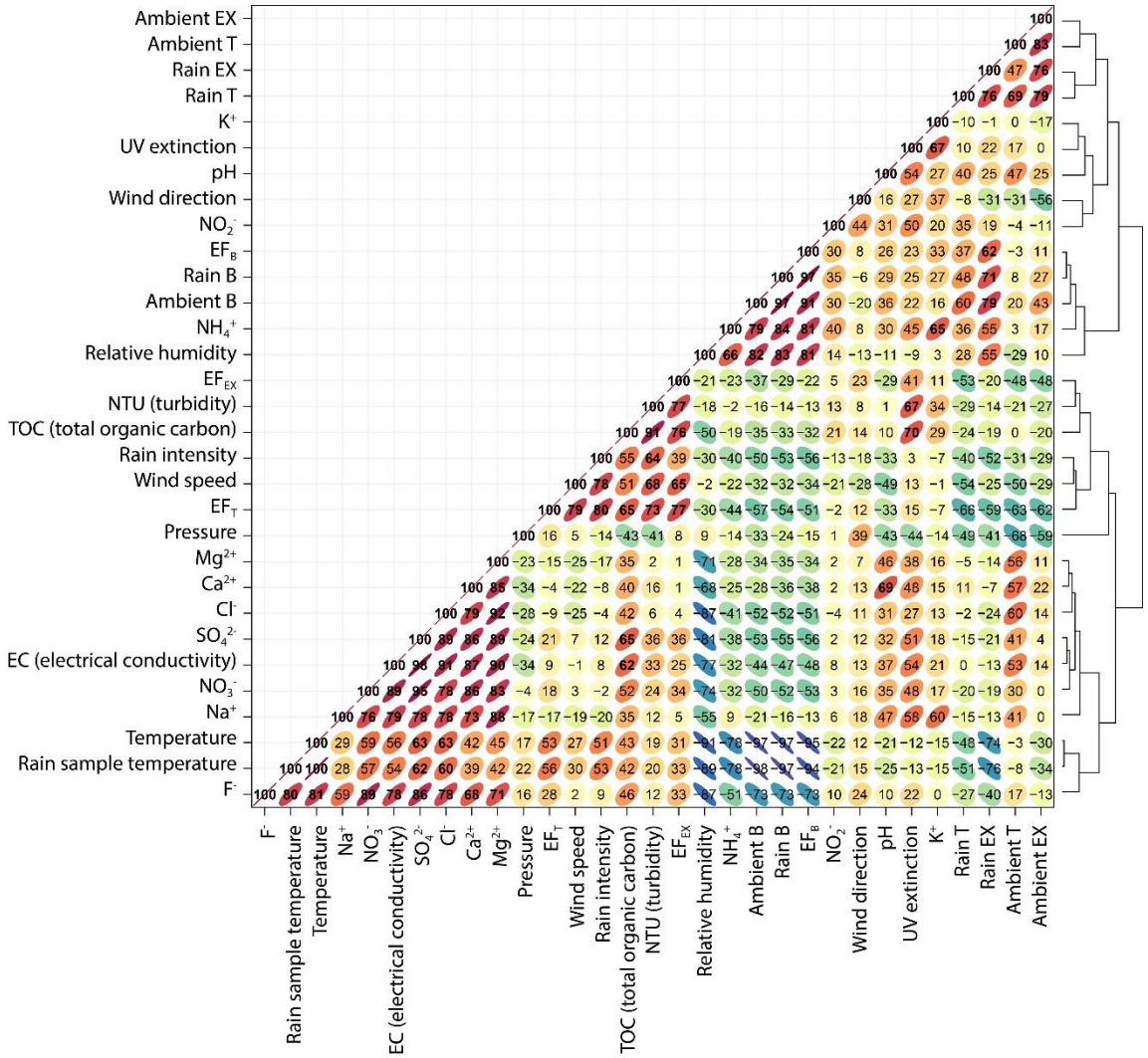

Figure S2. Correlation matrix.

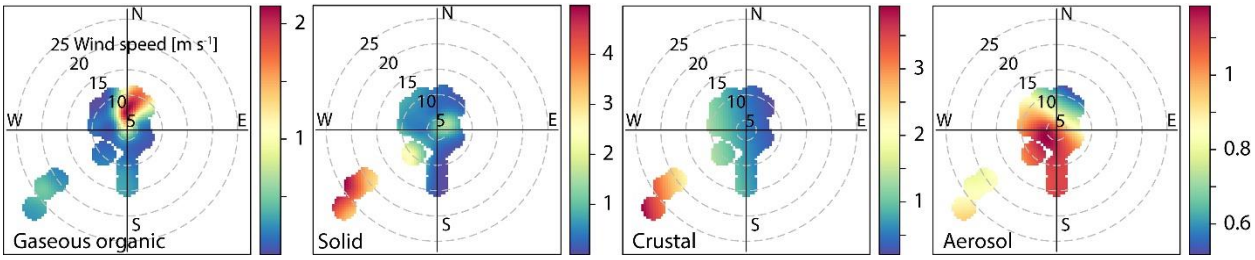

Figure S3. Source contribution (average = 1) dependence on wind components.

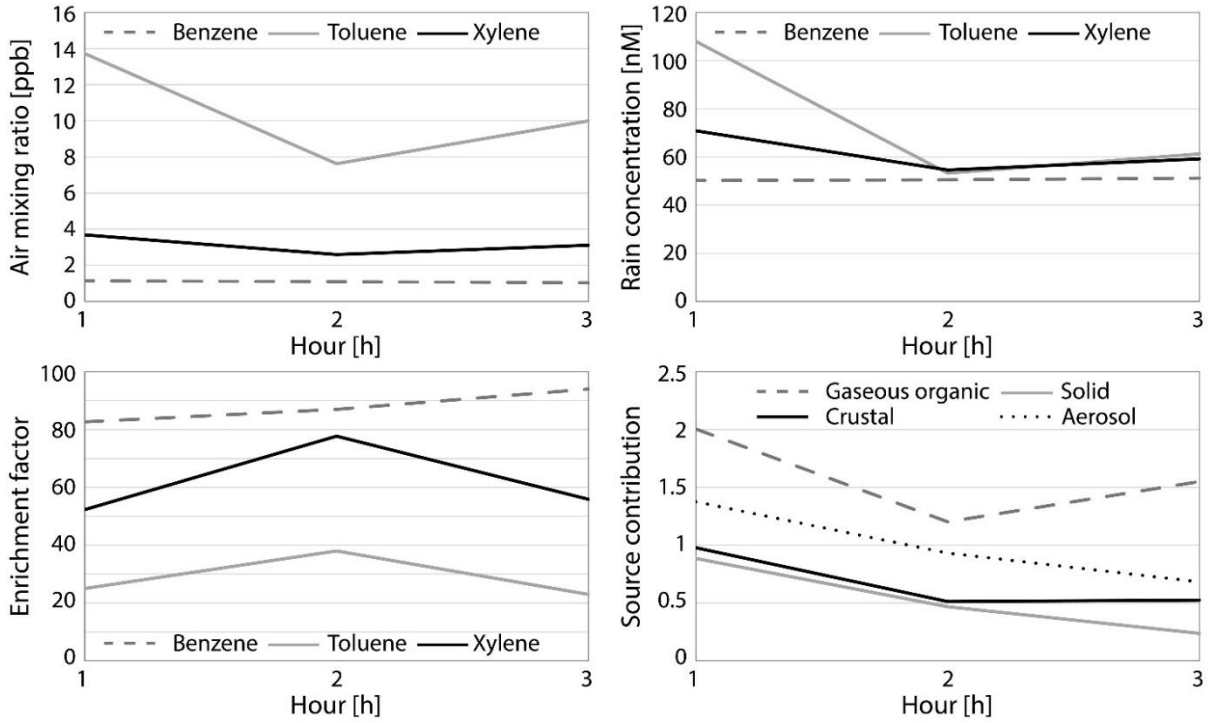

Figure S4. BTEX ambient mixing ratio and rain concentration, enrichment factor and source contribution dynamics during first three hours of the rainfall events.

Tables

Table S1. T-test sample degradation analysis: the first aliquot analyzed immediately after sampling, the second analyzed after all samples collected during each rain event.

Compound	Benzene		Toluene		Xylene and ethylbenzene	
	First aliquot	Second aliquot	First aliquot	Second aliquot	First aliquot	Second aliquot
Mean	0.372	0.370	0.555	0.548	0.619	0.609
Variance	0.109	0.109	0.146	0.154	0.046	0.065
Observations	2743	2743	2743	2743	2743	2743
df	5484		5480		5329	
t Stat	0.177		0.666		1.549	
P(T<=t) one-tail	0.430		0.253		0.061	
t Critical one-tail	1.645		1.645		1.645	
P(T<=t) two-tail	0.859		0.505		0.121	
t Critical two-tail	1.960		1.960		1.960	

Table S2. Descriptive statistics, detection limits (DL) and blank rain sample values for all measured and calculated parameters: rainwater chemical properties, ambient and rainwater BTEX concentrations and enrichment factors (EF), and meteorological parameters.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max	DL	Blank
Rain sample temperature [°C]	53	13.33	8.23	2.50	4.10	18.50	20.50	22.30	-10.0	/
NTU (turbidity) [NTU]	53	1.81	1.87	0.50	0.80	1.00	1.50	8.90	0.1	0.1
pH	53	6.01	1.00	3.70	5.20	6.20	6.60	8.20	2.0	6.96
UV extinction [m^{-1}]	53	3.44	2.56	0.70	1.60	2.60	5.00	12.00	0.1	0*
TOC (total organic carbon) [mg L^{-1}]	53	3.47	3.41	0.76	1.25	2.09	3.35	12.60	0.5	0.2
EC (el. conductivity) [µS cm^{-1}]	53	33.10	27.02	9.00	12.00	22.00	45.00	135.00	0.1	2.3
F^- [MEQ L^{-1}]	53	0.006	0.004	0.0027	0.001	0.006	0.008	0.015	0.0026	nd
Cl^- [MEQ L^{-1}]	53	0.027	0.046	0.0143	0.007	0.017	0.025	0.319	0.0143	0.0080
NO_2^- [MEQ L^{-1}]	53	0.0017	0.0015	0.0007	0.0011	0.0015	0.0020	0.0104	0.0004	nd
NO_3^- [MEQ L^{-1}]	53	0.0402	0.0252	0.0144	0.0210	0.0281	0.0516	0.1323	0.0081	0.0065
SO_4^{2-} [mg L^{-1}]	53	0.1073	0.1021	0.0146	0.0252	0.0792	0.1688	0.4917	0.0104	0.0080
Na^+ [MEQ L^{-1}]	53	0.0143	0.0287	0.0109	0.0052	0.0052	0.0052	0.1535	0.0109	0.0050
NH_4^+ [MEQ L^{-1}]	22	0.0433	0.0178	0.0183	0.0311	0.0406	0.0522	0.0978	0.0028	0.0025
K^+ [MEQ L^{-1}]	53	0.0074	0.0108	0.0026	0.0013	0.0038	0.0103	0.0705	0.0026	0.0007
Ca^{2+} [MEQ L^{-1}]	53	0.2420	0.1995	0.1000	0.0715	0.1735	0.3100	0.9330	0.1000	0.0290
Mg^{2+} [MEQ L^{-1}]	53	0.0250	0.0625	0.0208	0.0100	0.0100	0.0100	0.4500	0.0208	0.0030
Ambient B [ppbV]	22	1.06	0.15	0.75	0.96	1.12	1.15	1.22	0.08	/
Ambient T [ppbV]	53	8.81	7.70	0.54	1.99	4.33	15.82	24.91	0.12	/
Ambient EX [ppbV]	46	2.59	1.59	0.54	1.15	2.79	4.30	4.96	0.18	/
Rain B [nM]	22	50.28	10.15	37.22	44.38	48.85	52.52	82.64	10	2.50
Rain T [nM]	53	60.21	34.11	15.63	31.22	58.99	76.30	144.12	10	3.05
Rain EX [nM]	46	51.67	18.32	23.45	37.83	48.42	66.11	93.04	20	7.80
EF_B	22	90.73	22.46	61.00	74.50	80.50	111.00	131.00	1	/
EF_T	53	53.36	54.47	8.00	15.00	29.00	76.00	209.00	1	/
EF_{EX}	46	92.72	74.39	24.00	33.80	65.50	131.20	295.00	1	/
Temperature [°C]	53	14.25	9.22	3.20	3.40	19.50	22.20	24.90	-52.0	/
Relative humidity [%]	53	80.20	11.88	53.00	72.00	81.00	92.00	92.00	0.1	/
Wind speed [m s^{-1}]	53	8.54	4.90	3.00	6.00	7.00	9.00	27.00	0.1	/
Rain intensity [mm]	53	0.25	0.21	0.00	0.10	0.10	0.30	0.90	0.01	/
Pressure [mbar]	53	1011	3.86	1006	1008	1009	1014	1018	600	/

/ - not applicable; *Instrument zero set with ultrapure water; nd - not detected

Table S3. Unmix derived source composition.

Parameters	Gaseous organic	Solid	Crustal	Aerosol
NTU (turbidity)	8.8	79.7	11.5	0
UV extinction	33.3	46.4	7.2	13.1
TOC (total organic carbon)	0	71.6	0	28.4
EC (el. conductivity)	0	36.5	0	63.5
F^-	0	19.5	0	80.5
Cl^-	0	19.9	14.4	65.6
NO_3^-	0	31.3	0	68.7
SO_4^{2-}	0	42.2	0	57.8
Na^+	9.7	16.6	61.8	11.9
K^+	20.9	58.0	3.5	17.6
Ca^{2+}	15.1	20.2	0	64.7
Mg^{2+}	0	0	61.4	38.6
Rain B	99.0	0	1.0	0
Rain T	44.5	0.7	0	54.8
Rain EX	52.2	7.3	0	40.5
Average contribution	**18.9**	**30.0**	**10.7**	**40.4**

Table S4. Unmix derived source contribution statistics.

Parameters	r-Pearson	Mean Diff	RMSE	Slope	Intercept	r^2
NTU (turbidity)	0.96	-0.07	0.52	0.96	0.13	0.92
UV extinction	0.96	0.13	0.73	1.10	-0.48	0.92
TOC (total organic carbon)	0.94	0.26	1.18	0.94	-0.03	0.88
EC (el. conductivity)	0.92	9.18	10.72	1.01	-9.67	0.85
F^-	0.83	0.01	0.04	0.88	0.00	0.68
Cl^-	0.74	0.01	0.28	0.80	0.13	0.55
NO_3^-	0.85	0.13	0.83	0.92	0.09	0.72
SO_4^{2-}	0.95	2.12	1.60	1.03	-2.35	0.90
Na^+	0.95	0.00	0.06	1.00	0.00	0.91
K^+	0.80	0.01	0.15	1.05	-0.02	0.64
Ca^{2+}	0.74	0.40	2.70	1.35	-2.25	0.55
Mg^{2+}	0.96	0.04	0.06	0.99	-0.04	0.92
Rain B	0.97	3.95	5.95	1.02	-4.57	0.94
Rain T	0.77	4.83	21.91	0.92	0.67	0.60
Rain EX	0.84	1.83	13.26	0.95	0.73	0.70

Table S5. Unmix derived solution stability.

Parameter	Absolute source composition				Bootstrap percentiles															
	Gaseous organic	Solid	Crustal	Aerosol	Gaseous organic				Solid				Crustal				Aerosol			
					25th	50th	75th	95th	25th	50th	75th	95th	25th	50th	75th	95th	25th	50th	75th	95th
NTU	0.15(0.13)	1.4(0.6)	0(3)	0(3)	0.10	0.18	0.25	0.40	1.07	1.50	1.84	2.20	0.03	0.14	0.27	0.68	0	0.01	0.17	0.39
UV	1.2(0.4)	1.7(0.5)	0(3)	1(3)	0.79	1.10	1.33	1.78	1.34	1.71	2.06	2.53	0	0.16	0.35	0.75	0.22	0.42	0.61	0.91
TOC	0(0.3)	3(1)	0(6)	2(6)	0	0	0.17	0.42	2.20	3.08	3.72	4.50	0	0	0.29	1.40	0.25	0.72	1.12	1.67
EC	0(2)	15(6)	0(44)	3(43)	0	0	1.40	3.98	14.07	17.93	23.39	29.37	0	0	0	0	16.21	20.82	25.68	33.14
F^-	0(0.01)	0.02(0.01)	0(0.05)	4.00(0.05)	0	0	0.00	0.01	0.03	0.04	0.05	0.06	0	0.00	0.01	0.03	0.06	0.07	0.08	0.11
Cl^-	0(0.06)	0.1(0.1)	0.1(0.5)	5.0(0.4)	0	0	0.02	0.06	0.15	0.20	0.29	0.39	0.06	0.11	0.19	0.38	0.22	0.31	0.42	0.60
NO_3^-	0(0.2)	0.8(0.4)	0(1.3)	6.0(1.4)	0.03	0.15	0.28	0.50	0.80	1.11	1.45	1.76	0	0	0	0.40	1.09	1.33	1.61	2.26
SO_4^{2+}	0(0.5)	3(1)	0(4)	7(4)	0	0	-0.27	0.07	2.71	3.80	4.56	5.70	0	0	0	0.18	2.58	3.17	3.81	4.62
Na^+	0.02(0.01)	0.03(0.02)	0.13(0.06)	8.00(0.04)	0.01	0.02	0.03	0.04	0.02	0.04	0.05	0.08	0.09	0.12	0.15	0.24	0.01	0.02	0.03	0.05
K^+	0.05(0.04)	0.15(0.05)	0(1)	9(1)	0.02	0.03	0.06	0.11	0.11	0.15	0.17	0.22	0	0.01	0.03	0.10	0.00	0.04	0.08	0.14
Ca^{2+}	0.8(0.6)	1(1)	0(18)	10(18)	0.49	0.80	1.13	1.98	0.78	1.37	2.01	2.70	0	0	0.58	2.57	1.75	2.81	3.68	5.56
Mg^{2+}	0(0.02)	0(0.03)	0.15(0.13)	11.0(0.1)	0	0.00	0.02	0.04	0	0	0.01	0.04	0.09	0.13	0.17	0.24	0.05	0.08	0.10	0.14
Rain B	26(8)	0(2)	0(2)	12(2)	18.13	22.21	28.24	37.14	0	0	0.38	1.13	0	0.23	1.22	3.54	0	0	-0.91	0.48
Rain T	10(10)	0 (6)	0(24)	13(25)	18.84	28.15	34.31	43.29	1.20	5.90	10.27	16.78	0	0	1.28	12.31	17.74	25.07	32.58	45.98
Rain EX	24(8)	3(4)	0(37)	14(36)	17.57	22.36	28.48	35.99	4.40	6.79	10.04	13.50	0	0	2.36	11.09	8.75	14.90	19.69	24.44

* Standard deviations are shown in parenthesis.

Table S6. Enrichment factors and ambient BTEX mixing ratios (ppb) for summer and autumn; and for low (<5 m s^{-1}) and high (≥5 m s^{-1}) wind speed events.

	Benzene	Toluene	Ethylbenzene and xylene	EF_B	EF_T	EF_{EX}
Summer		6.67	1.97		98.54	166.55
Autumn	1.05	11.51	3.19	223.59	47.04	146.62
Total	1.05	8.73	2.56	223.59	76.61	156.79
Low wind speed	0.98	16.71	3.33	227.85	41.35	126.15
High wind speed	1.06	6.69	2.33	222.70	85.63	166.16

Table S7. Mean variable and Unmix-derived-source-contribution importance for BTEX enrichment factor prediction obtained by guided regularized random forest (importance standard deviation is shown in parentheses).

Parameter	Importance		
	Benzene	Toluene	Ethylbenzene and xylene
Ca^{2+}	43(4)	600(40)	1800(140)
Cl^-	1700(400)	970(160)	2400(600)
EC (electrical conductivity)	110(20)	900(200)	8500(1600)
F^-	500(200)	400(60)	4700(700)
K^+	40(10)	380(60)	1700(400)
Ambient EX	2600(200)	6000(2000)	11500(400)
Rain EX	1500(300)	2400(300)	7400(200)
Ambient B	15600(800)	700(150)	2300(200)
Rain B	15800(800)	820(120)	1490(180)
Ambient T	68(80)	49000(5000)	60000(10000)
Rain T	128(15)	6200(1900)	20900(1200)
Mg^{2+}	0(1)	140(20)	770(140)
Na^+	6(2)	120(20)	1300(300)
NH_4^+	15200(800)	60(30)	530(80)
NO_2^-	50(20)	450(50)	1300(200)
NO_3^-	130(40)	600(300)	19100(1500)
NTU (turbidity)	22(6)	30200(1200)	36000(2000)
Pressure	630(50)	14000(1000)	17600(1400)
pH	60(20)	290(40)	3300(400)
Rain intensity	16(11)	600(300)	3700(900)
Rh (relative humidity)	15100(800)	270(80)	1160(170)
SO_4^{2-}	130(60)	1600(300)	15000(1900)
Temperature	22400(1600)	570(110)	2700(300)
TOC (total organic carbon)	100(10)	28400(1100)	40000(2000)
Rain sample temperature	20300(1400)	813(12)	7500(700)
UV extinction	50(20)	1600(200)	4400(1100)
Wind speed	53(9)	1200(300)	7200(1000)
Gaseous organic	2014(40)	38600(600)	65000(1100)
Solid	2100(40)	56000(900)	81700(1500)
Crustal	2450(50)	28100(600)	51700(1000)
Aerosol	2710(50)	19400(400)	31100(600)

References

Aha, D.W., Kibler, D., Albert, M.K., 1991. Instance-based learning algorithms. Mach. Learn. 6, 37-66.

Breiman, L. 2001. Random forests. Mach. Learn. 45, 5-32.

Cleary, J.G., Trigg, L.E., 1995. K*: An instance-based learner using an entropic distance measure. In: Proceedings of the 12th International Conference on Machine learning, 108-114.

Cohen, W.W., 1995. Fast effective rule induction. In: Proceedings of the 12th international conference on machine learning, 115-123.

Frank, E., Mayo, M., Kramer, S. 2015, April. Alternating model trees. In: Proceedings of the 30th Annual ACM Symposium on Applied Computing, 871-878. ACM.

Frank, E. 2014. Fully supervised training of Gaussian radial basis function networks in WEKA. In: Computer Science Working Papers. Hamilton, NZ: Department of Computer Science, The University of Waikato.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I.H., Trigg, L., 2005. Weka. In: Data Mining and Knowledge Discovery Handbook, 1305-1314. Springer US.

Frank, E., Hall, M., Pfahringer, B., 2003. Locally Weighted Naive Bayes. In: 19th Conference in Uncertainty in Artificial Intelligence, 249-256.

Graczyk, M., Lasota, T., Trawiński, B., 2009. Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA. In: International Conference on Computational Collective Intelligence, 800-812. Springer Berlin Heidelberg.

Haykin, S., 1994. Neural Networks: A Comprehensive Foundation, 1st Ed., Upper Saddle River, NJ, USA: Prentice Hall PTR

IBkLG(2015). Instance Based-kNN Log and Gaussian [Computer software]. Available at https://github.com/sheshas/IBkLG.

Kalmegh, S. 2015. Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and Random Tree for Classification of Indian News. Int. J. Innov. Sci. Eng. Technol. 2, 438-446.

Rasmussen, C. E. (2006). Gaussian processes for machine learning.

Rousseeuw, P.J., Leroy, A.M., 2005. Robust regression and outlier detection. John Wiley & Sons.

Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K., 1999. Murthy: Improvements to the SMO Algorithm for SVM Regression. In: IEEE Transactions on Neural Networks.

Shi, P., Tsai, C.L., 2002. Regression model selection - a residual likelihood approach. J.R. Statist. Soc. B, 64, 237-252.

Wang, Y., Witten, I.H. 1996. Induction of model trees for predicting continuous classes. In Proceedings of Poster Papers, Ninth European Conference on Machine Learning, 1997

Wang, Y., & Witten, I. H. (2002). Modeling for optimal probability prediction.

Weka Sourceforge, 2016. http://weka.sourceforge.net/packageMetaData Accessed: June 2016.

Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. Adv. Space Res. 41, 1955-1959.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. J. Comput. Graph. Stat., 15, 265-286.