

Comparison of machine learning approaches with a general linear model to predict personal exposure to benzene

Aquilina, Noel J.; Delgado Saborit, Juana Maria; Bugelli, Stefano ; Padovani Ginies, Jason; Harrison, Roy

DOI:

[10.1021/acs.est.8b03328](https://doi.org/10.1021/acs.est.8b03328)

License:

Other (please specify with Rights Statement)

Document Version

Peer reviewed version

Citation for published version (Harvard):

Aquilina, NJ, Delgado Saborit, JM, Bugelli, S, Padovani Ginies, J & Harrison, R 2018, 'Comparison of machine learning approaches with a general linear model to predict personal exposure to benzene', *Environmental Science and Technology*. <https://doi.org/10.1021/acs.est.8b03328>

[Link to publication on Research at Birmingham portal](#)

Publisher Rights Statement:

Checked for eligibility: 26/09/2018

This document is the Accepted Manuscript version of a Published Work that appeared in final form in *Environmental Science and Technology*, copyright © American Chemical Society after peer review and technical editing by the publisher.

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

1
2
3
4
5
6
7
8
9
10

COMPARISON OF MACHINE LEARNING APPROACHES WITH A GENERAL LINEAR MODEL TO PREDICT PERSONAL EXPOSURE TO BENZENE

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

**Noel J. Aquilina^{1,2}, Juana Maria Delgado-Saborit^{1,‡},
Stefano Bugelli³, Jason Padovani Ginies³ and
Roy M. Harrison^{1,†},**

**¹Division of Environmental Health and Risk Management
School of Geography, Earth and Environmental Sciences University of
Birmingham
Edgbaston, Birmingham B15 2TT
United Kingdom**

**²Department of Geosciences
Faculty of Science, University of Malta
Msida MSD 2080, Malta**

**³Department of Physics
Faculty of Science, University of Malta
Msida MSD 2080, Malta**

* To whom correspondence should be addressed
Tele: +44 121 414 3494; Email: r.m.harrison@bham.ac.uk

†Also at: Department of Environmental Sciences / Center of Excellence in Environmental Studies, King Abdulaziz University, PO Box 80203, Jeddah, 21589, Saudi Arabia

‡ Now at: ISGlobal, Barcelona Institute for Global Health - Campus MAR, Barcelona Biomedical Research Park (PRBB), Doctor Aiguader, 88, 08003 Barcelona, Spain

26 **ABSTRACT**

27 Machine Learning Techniques (MLTs) offer great power in analysing complex datasets and have not
28 previously been applied to non-occupational pollutant exposure. MLT models that can predict personal
29 exposure to benzene have been developed and compared with a standard model using a linear regression
30 approach (GLM). The models were tested against independent datasets obtained from three personal
31 exposure measurement campaigns. A Correlation-based Feature Subset (CFS) selection algorithm
32 identified a reduced attribute set, with common attributes grouped under the use of paints in homes;
33 upholstery materials; space heating and environmental tobacco smoke as the attributes suitable to predict
34 the personal exposure to benzene. Personal exposure was categorised as low, medium and high, and for
35 big datasets, both the GLM and MLTs show high variability in performance to correctly classify >90%ile
36 concentrations, but the MLT models have a higher score when accounting for divergence of incorrectly
37 classified cases. Overall, the MLTs perform at least as well as the GLM and avoid the need to input
38 microenvironment concentrations.

39

40 **Keywords:** Benzene; personal exposure; machine learning techniques; general linear model;
41 dimension reduction

42

43 **1. INTRODUCTION**

44 Exposure assessment is an important analytical tool for evaluating the likelihood and extent of actual or
45 potential exposure of people to pollutants and is an important component of any health risk assessment
46 and epidemiological study. Exposure to chemicals from environmental and occupational settings can be
47 characterized in different ways¹. Direct methods such as personal monitoring and biomarkers are
48 considered to be accurate for exposure assessment yet are costly to study big populations. Indirect
49 information gained through questionnaires and diaries accompanied by environmental monitoring can be
50 used to develop exposure models. Modelling techniques have greatly improved the assessments and are
51 likely to be important in future studies since direct measurement of exposure is often too expensive and
52 time consuming.

53

54 In recent years, exposure assessment to atmospheric pollutants has been conducted mainly either by
55 deterministic methods, strengthened by geographical information systems and geostatistical techniques²,
56 or by a statistical approach³. In the last 20 years statistical approaches have focused on regression
57 techniques and source apportionment while probabilistic modelling was mainly done by Monte Carlo
58 analyses and Bayesian statistics. The main criticisms of many exposure assessments have been a reliance
59 on overly conservative assumptions about exposure, as well as the problem of how to model properly the
60 highly exposed populations that generally are small in number^{4,5}. The earlier published work has shown
61 a limited ability of methods based upon measurement of microenvironment concentrations to provide an
62 accurate quantitative reconstruction of personal exposure (PE). This is no doubt due to the variability in
63 concentrations within a given type of microenvironment and poorly quantified contributions from
64 sporadic sources. Since machine learning techniques (MLTs) function without *a priori* assumptions of
65 pathways and have great power to extract meaningful patterns and trends from datasets, we have for the
66 first time applied MLTs to the modelling of non-occupational PE to a key air pollutant, benzene.

67 Ideally a PE model should be able to predict the degree of exposure of an individual based on a minimum
68 number of input attributes. The model for benzene developed by Delgado-Saborit et al.⁶ predicted the
69 PE by integrating the time fraction spent in each microenvironment times the concentration of benzene
70 in the microenvironment visited, and also accounted for external factors that might affect exposure as
71 add-on variables, using a linear regression approach. The best model that was able to predict PE with
72 independence of measurements was based upon certain time-activity attributes. Other studies conducted
73 by Heavner et al.⁷, Austin et al.⁸, Ilgen et al.⁹, Yang et al.¹⁰, Edwards et al.¹¹, Batterman et al.¹², Curren
74 et al.¹³, Zuraimi et al.¹⁴ and Song et al.¹⁵, through source apportionment, have identified sources of
75 benzene that were consistent with the variables that were introduced in the above-mentioned model. The
76 model identified the most important non-weather-related variables for benzene exposures, highlighting
77 the influence of personal activities, use of solvents, and exposure to environmental tobacco smoke (ETS)
78 on PE levels.

79
80 MLTs are used for several air quality applications, including forecasting of airborne pollutants such as
81 PM_{2.5} levels¹⁶, PM₁₀ levels^{17,18,19,20,21,22}, SO₂, CO and NO and NO₂ and O₃^{19,23}, and particle-phase PAH²⁴.
82 One study uses a MLT to model benzene exposures, but in an occupational setting²⁵.

83
84 In this study, MLT models were trained and tested on benzene PE data that was collected during three
85 PE campaigns, namely; MATCH²⁶, TEACH²⁷ and EXPOLIS²⁸. The performance of the MLT models in
86 classifying personal exposures was tested and results are discussed in the light of their usefulness for risk
87 assessment and epidemiological studies.

88 89 **2. METHODOLOGY**

90 **2.1 Description of Datasets**

91 Three datasets were employed in training and testing the models using MLTs. These datasets as described

92 in detail below were the MATCH, the EXPOLIS and the TEACH databases. Descriptive statistics appear
93 in Table S1 and Figure S3.

94

95 The MATCH (**M**easurement and **M**odelling of **A**ir **T**oxics **C**oncentrations for **H**ealth **S**tudies) study's
96 main objective was to optimize a model of PE based on microenvironment concentrations and
97 time/activity diaries and to compare the modelled with measured exposures in an independent dataset⁶.

98 The subjects for this study, enrolled to measure their PE to a suite of air toxics were recruited based upon
99 a set of inclusion determinants that affected exposure, namely: location, living in houses with heavy
100 trafficked roads (termed as first line houses), having a house with an integral garage, and exposure to
101 ETS²⁶. PE of 100 adult non-smokers living in three UK locations, namely London, West Midlands, and
102 rural South Wales, to 15 VOCs was measured using an actively pumped sampler carried around by the
103 subjects for five consecutive 24 hr periods, following their normal lifestyle.

104

105 The EXPOLIS (Air Pollution Exposure Distributions within Adult Urban Populations in Europe) study
106 focused on adults living in cities in seven European countries (Helsinki, Athens, Basel, Grenoble, Milan,
107 Prague, Oxford), exposed to air pollutants in their homes, workplaces and other common urban micro-
108 environments²⁷ from 1996-1998. The 401 subjects who participated in this study were chosen according
109 to certain criteria which are found in the EXPOLIS manual²⁷. This study was based on a single 48 hr
110 sampling period using a suitcase containing the sampler.

111

112 The TEACH (**T**oxic **E**xposure **A**ssessment, a **C**olumbia / **H**arvard) study was designed to characterize
113 levels and factors of PE to urban air toxics among high school students in Los Angeles and New York
114 from 1999-2000²⁸. This study involved 87 students who carried a backpack for 48 hr over two different
115 sampling periods, one in summer and another in winter.

116

117 In the three studies the number of samples represented either a 24 hr or 48 hr PE sampling. If the subjects
118 were monitored for several days, each sample is treated separately and not pooled per subject. In the
119 three studies the subjects filled questionnaires collecting information about subject demographics,
120 lifestyle, home description, products stored within the house, activities performed, places visited,
121 ventilation, and ETS presence, as described in detail elsewhere²⁹. The questionnaires were different for
122 the three studies but most of the information gathered was similar. These questionnaires may be referred
123 to in Harrison et al.²⁹ for MATCH, Kinney et al.³⁰ for TEACH and Hanninen et al.²⁷ for EXPOLIS.

124

125 **2.2 Attribute Selection for dimension reduction**

126 Attribute subset selectors are a collection of algorithms that try to find and remove irrelevant and
127 redundant attributes³¹, an exercise termed as dimension reduction that is required in generating robust
128 PE models requiring a minimal number of attributes.

129

130 Therefore, the initial stage before the model could be built requires dimension reduction, where a number
131 of variables that affect/predict most of the measured level of benzene exposure for a given compound
132 were chosen. Dimension reduction attempts to identify and remove those features which increase
133 computation time, but not model performance. In this study a Correlation-based Feature Subset (CFS)
134 selection algorithm was used. Further information on this algorithm can be found in the Supporting
135 Information.

136

137 **3. GENERAL LINEAR MODELLING TO MODEL PE TO BENZENE**

138 A more common approach to modelling PE is by using a General Linear Model (GLM) which was used
139 in various studies, such as to model the effect of VOCs exposure during pregnancy to newborn's birth
140 weight³², to find the relationship between PE to VOCs and home, work and outdoor concentrations³³, to

141 evaluate vehicle exposure to certain VOCs including benzene in urban areas³⁴. In this study a GLM was
142 developed and compared with the MLTs described in Section 4.

143

144 The GLM is a combination of two major model types, namely regression models and analysis of variance
145 models. For this study, where only one dependent (continuous) variable was available, GLMs were used.
146 Here, all the attributes were included into the model and the least significant was removed manually one
147 at a time. This process was repeated until the remaining variables left were all statistically significant
148 ($p < 0.05$). This was also used in previous exposure studies such as benzene exposure³⁵ and exposure to
149 ETS³⁶.

150

151 Since benzene concentration is a continuous variable, the Poisson and Binomial distributions are not
152 suitable to model such data, thus Gaussian, Gamma and Inverse Gaussian distributions were fitted. The
153 GLMs with the lowest Akaike information criteria and Bayesian information criteria were applied for
154 the three studies and further details are given in the Supporting Information and Table S2.

155

156 **4. MACHINE LEARNING TECHNIQUES TO MODEL PE TO BENZENE**

157 Our earlier research⁶ was based upon the use of simple additive models in which microenvironment
158 concentrations were summed in a time-weighted manner, or multiple linear regression approaches in
159 which key influences upon exposure were identified and added in weighted manner to obtain the best
160 overall fit to the measured exposures. Such methods require *a priori* assumptions as to the most
161 important factors/sources influencing exposure and assume that total exposure is the linear sum of a range
162 of weighted contributions.

163

164 MLTs used in this study are computer-based algorithms which recognise features in datasets which when
165 combined give a good fit to an outcome variable, in this case the measured PE. The algorithms learn

166 directly from the data and improve their performance as they are provided with more samples. MLTs
167 can be either supervised or unsupervised. In the former case, a known set of input data and output
168 responses is used to combine input variables in such a way as to predict the outcome using classification
169 or regression methods. In the unsupervised learning case, methods such as clustering are used to
170 recognise patterns in the data without reference to the outputs.

171

172 In several applications predictions have been aided by the application of MLTs³⁷. Algorithms are
173 generally trained with previously available data and allow predictions in the testing phase³⁸. The success
174 of an analysis can thus be defined as the ability of such algorithms to predict the correct status of unseen
175 data.

176

177 In the realm of PE to atmospheric pollutants, accuracy of classification strategies can be affected
178 negatively with the use of too many features in the classification. This may lead to overfitting, in which
179 noise or irrelevant features may decrease classification accuracy because of the finite size of the training
180 samples³⁹. The mining workbench program used for developing the MLT models was the Waikato
181 Environment for Knowledge Analysis (WEKA)^{40,41}. Further information on the MLTs used in this
182 research is given in the Supplementary Information.

183

184 After redundant attributes were removed and a Reduced Attribute Set (RAS) had been selected, for the
185 datasets available and the application presented the DT, NNGE, KStar, ANN and RF algorithms were
186 chosen for machine learning using their standard settings in WEKA.

187

188 **5. MODELS AND CLASSIFICATION OF EXPOSURE**

189 Using WEKA the models were trained on a randomly chosen 75% of the dataset and validated using the
190 remaining 25%. A 10-fold cross validation was also carried out.

191 To have a consistent method across the three studies considered rather than one based on various
 192 legislative/directive limits or guideline values that serve for policy making purposes, benzene
 193 concentrations were categorised as Low (L), Medium (M) and High (H) based on 10-90%iles and 30-
 194 70%iles and 30-90%iles as summarised in Table 1 in order to evaluate the robustness of the different
 195 models used in correctly classifying the PE range.

196

197 The five MLTs and the GLM were run using the RAS for the testing dataset (25% of the unseen dataset)
 198 based on the classification bins defined in Table 1.

199

200 **Table 1:** The bin limit values for benzene (in $\mu\text{g m}^{-3}$) determined by the 10%ile and 90%ile, 30%ile and
 201 70%ile and the 30%ile and 90%ile percentiles.

202

Study	Low (L)		Medium (M)			High (H)	
	10%ile	30%ile	10-90%ile	30-70%ile	30-90%ile	70%ile	90%ile
MATCH	< 0.7	< 1.0	0.7 – 3.5	1.0 – 2.0	1.0 – 3.5	> 2.0	> 3.5
EXPOLIS	< 0.8	< 2.4	0.8 – 13.0	2.4 – 6.0	2.4 – 13.0	> 6.0	> 13.0
TEACH	< 1.8	< 2.8	1.8 – 7.3	2.8 – 4.8	2.8 – 7.3	> 4.8	> 7.3

203

204 6. RESULTS

205 6.1 Testing Attribute Selection and Accuracy of Classification

206 ACFS algorithm was used to remove irrelevant and redundant variables from a Full Attribute Set (FAS).

207 A RAS for each study was obtained and the important attributes identified by CFS were compared with

208 similar attributes identified in other studies and are summarized in Table 2.

209 **Table 2:** Reduced number of attributes (RAS) using the CFS algorithm, which are able to predict the
 210 continuous benzene concentration for (a) MATCH, (b) EXPOLIS, (c) TEACH.

211

(a) MATCH	
Variable	Reference supporting variable
Gardening products used	
Visited hospital	Delgado-Saborit et al. ⁶
Visited petrol station	Wallace ⁴²
Using subway	Delgado-Saborit et al. ⁶

Being in presence of someone painting	Delgado-Saborit et al. ⁶
Rubber-backed nylon carpets laid in house	
Keeping car in garage	Batterman et al. ¹²
Storing paints in garage	Delgado-Saborit et al. ⁶
Time spent at constant ETS	Heavner et al. ⁷
Gas and other heating used	Delgado-Saborit et al. ⁶
Urban location	Delgado-Saborit et al. ⁶

212

(b) EXPOLIS	
Variable	Reference supporting variable
Visited gas station	Wallace ⁴²
Used chemicals and glues	Wallace ⁴²
Having carpets other than wall to wall	
Having double glazing windows & chipboard	
Room height	
Having water damage	
Keeping pets in the house	
Smoking in the house	Edwards et al. ¹¹
Amount of heavy traffic passing in front of home	Wallace ⁴²
Using district heating	
Use gas for cooking	

213

(c) TEACH	
Variable	Reference supporting variable
Smoking	Edwards et al. ¹¹
Having a door leading to garage	Batterman et al. ⁴³
Having a diesel car in garage	Batterman et al. ⁴³
Having curtains, Upholstering furniture, double glazing	
Plaster, chipboards or paper walls	
Painted walls	Song et al. ¹⁵
Season	
Glue was used	Wallace ⁴²
City	Delgado-Saborit et al. ⁶
Fireplace or a stove was used for heating	
Water damage	

214

215 In order to assess the performance of the MLTs, these were run using the FAS and the RAS from the
 216 three studies, where the RAS was obtained by CFS as explained above. Table S3 summarizes the overall

217 accuracy obtained for predicting PE to benzene when using the FAS and the RAS for classification.

218

219 The overall performance of the MLTs in a 10-fold cross validation and a 25% testing dataset using a 75%

220 training dataset for classification determined by 10 and 90 percentiles using the RAS are presented in

221 Tables S4 and S5 respectively. The accuracy for the MLTs was calculated via a confusion matrix

222 available in WEKA that was generated in order to compare the various models used in trying to predict

223 PE (Supporting Information, Table S6). The matrix, for each model used, summarizes the correctly

224 classified instances and also indicates in which category the model wrongly classified instances when

225 compared to the corresponding measured instances. The degree of accuracy of the models can then be

226 determined by calculating the percentage of instances correctly classified and attributed to the correct

227 concentration range bin. Table S7 compares the performance of the MLTs with the GLM in correctly

228 classifying the exposure classes.

229

230 If these models are to be used for epidemiology or risk assessment applications, the need for correct

231 classification of the PE in different exposure categories varies according to the choice of the percentile

232 ranges chosen in this paper (10-90, 30-70 and 30-90%iles). A point ranking system (Table 3) has been

233 devised for the abovementioned applications and applied to the confusion matrix (Table S6) in order to

234 identify which model scores best in classifying the modelled concentrations in the correct classification

235 categories (L, M and H) as the corresponding measured concentrations. Table 4 shows the total ranking

236 of each model based on the point ranking system summarised in Table 3.

237

238 The scoring scheme for epidemiology applications penalised extreme misclassification highly (i.e. H to

239 L and L to H), and lesser misclassification less harshly with incorrect prediction of M as L or H losing

240 more points than the reverse error. The rationale was that epidemiology depends heavily upon a gradient

241 of exposures in which the H and L are most important in defining the distribution.

242 **Table 3:** Point ranking system devised for our models if they are to be used in epidemiology and risk
 243 assessment applications to predict benzene correctly in three studies.
 244

Epidemiology applications	
Accuracy of Classification	Ranking Points
Correct classification	No. of instances \times (+1 point)
Incorrect classification (H as L or L as H)	No. of instances \times (-3 points)
Incorrect classification (M as H or as L)	No. of instances \times (-2 points)
Incorrect classification (L or H as M)	No. of instances \times (-1 point)
Risk Assessment applications	
Accuracy of Classification	Ranking Points
Correct classification	No. of instances \times (+1 point)
Incorrect classification (H as L)	No. of instances \times (-5 points)
Incorrect classification (H as M)	No. of instances \times (-4 points)
Incorrect classification (L as H)	No. of instances \times (-3 points)
Incorrect classification (M as H or L)	No. of instances \times (-2 points)
Incorrect classification (L as M)	No. of instances \times (-1 point)

245

246
247
248
249

Table 4: Ranking of the different models' performance to predict benzene correctly in three studies. Numbers in bold indicate the models which ranked highest in correctly classifying instances in L, M and H exposure categories.

APPLICATION	MODEL	Study								
		MATCH			EXPOLIS			TEACH		
		10-90%iles	30-70%iles	30-90%iles	10-90%iles	30-70%iles	30-90%iles	10-90%iles	30-70%iles	30-90%iles
Epidemiology	DT	55	-15	7	56	-16	20	20	-12	2
	RF	61	-10	19	55	6	20	20	-30	5
	ANN	56	-5	17	44	-21	3	14	-18	-4
	NNGE	56	-32	-3	44	-52	-21	14	-20	4
	KStar	61	-9	18	46	-23	-5	4	-36	-21
	GLM	61	-1	34	41	1	8	32	24	26
Risk Assessment	DT	43	-60	-5	35	-61	-1	8	-34	-10
	RF	49	-46	7	45	-22	5	8	-58	-7
	ANN	50	-43	11	26	-57	-15	2	-41	-13
	NNGE	50	-84	-9	29	-72	-32	2	-49	-4
	KStar	49	-65	12	31	-61	-17	-4	-66	-28
	GLM	52	-37	25	20	-11	-13	32	21	26

250
251

252 The scoring system for risk assessment applications penalised extreme misclassification at the higher end
253 highly (i.e. H to L), with a decreasing degree of penalization as follows: incorrect prediction of H as M
254 > incorrect classification from the lower end to the higher end, followed by incorrect prediction of M as
255 L or H. Classifying incorrectly L cases in the M bin was the least harshly penalised.

256

257 The rationale was related to one of the aims of risk assessment, which is to identify those cases exposed
258 to high concentrations of benzene that would require subsequent actions to reduce their exposure.
259 However, if the model fails to identify the highly exposed subjects (e.g. H case classified as M or L),
260 these cases will continue to be exposed to high concentrations of benzene without acknowledging the
261 need of exposure reduction actions. Equally if a subject is not exposed to benzene, but the model
262 classifies the case as a high exposed subject, this will trigger actions to reduce his/her exposure, which
263 might incur an economic cost and/or disruption of the subject activities in order to reduce the benzene
264 exposure that initially are not required.

265

266 Table 4 shows that overall, the GLM performs better than the MLTs. For MATCH, KStar, RF and GLM
267 would be more suitable for epidemiology applications for the 10-90%iles categorisation, while the GLM
268 performs better for the 30-90%ile categorisation. However, for risk assessment applications, if the 10-
269 90%iles categorisation is used all MLTs perform approximately in the same way as the GLM, whilst the
270 latter model would be more suitable while for the 30-90%iles categorisation. For EXPOLIS, irrespective
271 of categorisation, RF and DT would be more suitable for epidemiology applications, while RF would be
272 more suitable for risk assessment applications. For TEACH the situation is clearer, for any exposure
273 categorisation and for both epidemiology applications and risk assessment applications the GLM
274 outperformed any MLT in predicting PE. For small datasets such as TEACH it appears none of the MLTs
275 seem satisfactory. For the more demanding 30-70%ile dataset, the GLM consistently outperforms the
276 MLTs.

277 The percentages of correctly classified instances per exposure category, for each study considered are
278 presented in Supporting Information Table S7. One can note that for predicting H exposures, the GLM
279 is better than MLTs when the dataset is small. When using a 30-70%ile classification (see Table S7), for
280 TEACH, DT and ANN perform equally well as the GLM. For larger datasets like EXPOLIS, using any
281 exposure categorisation, GLM outperforms MLT correctly classifying 87-100% of the instances. For
282 MATCH for predicting H exposures, using any categorisation, ANN, NNGE, KStar and the GLM can
283 correctly predict 63% of the instances. If a 30-70%ile categorisation is used, the GLM outperforms all
284 MLTs

285

286 To supplement the prediction based on a 75%-25% split (Table S4), a 10-fold cross validation was
287 performed with the three datasets, whose results are presented in Table S5. If one views the overall
288 performance of the MLTs for the 10-90%ile and the 30-90%ile classification using the RAS, they are
289 somewhat similar to those obtained in Table S4. The Kappa statistic, the Mean Absolute Error (MAE)
290 and the Root Mean Square Error (RMSE) in Tables S4 and S5 indicate there is a greater variance in the
291 individual errors in the dataset. However, if one focuses on the prediction of the H exposure using the
292 10-90%ile categorisation, based on the area under the Receiver Operating Curve (ROC) and the F-
293 Measure presented in Table S4, RF shows the better performance for the three studies. KStar performs
294 equally well in MATCH. For TEACH, the MLTs perform similarly with RF appears to be the best
295 candidate for small datasets. From Tables S4 and S5, in EXPOLIS, the best MLT to predict H exposures
296 using a 30-90 categorisation would be RF, for MATCH they would be KStar and RF while for TEACH,
297 although the performance of MLTs is not appreciable, ANN and RF still appear to perform better.

298

299 While the majority of the MLTs predict only exposure category, two of the MLTs (KStar and ANN) and
300 the GLM were able to predict also continuous data. The R^2 value and the Predicted vs Measured gradient
301 are shown in Table 5. DT, NNGE and RF are not included as they do not give R^2 values for direct

302 comparison with the GLM. This table indicates that the performance of the model is not determined
303 solely by the R² value; in fact, the predicted: measured ratio indicates that the GLM perform better in
304 predicting a PE value closer to the measured values when compared to the MLTs, at least in the studies
305 considered.

306
307
308

Table 5: Predicting continuous data results for benzene.

Study	Model	Predicted : Measured Ratio	R²
MATCH	KStar	0.669	0.321
	ANN	0.728	0.410
	GLM	1.004	0.390
EXPOLIS	KStar	0.651	0.302
	ANN	0.237	0.004
	GLM	1.021	0.240
TEACH	KStar	0.031	0.001
	ANN	1.579	0.472
	GLM	1.000	0.970

309

310 7. DISCUSSION

311 This study presents several PE models developed using different MLTs using benzene PE data collected
312 during three independent PE campaigns, namely; MATCH²⁶, and EXPOLIS²⁷ and TEACH²⁸. The first
313 step in the model development was to select those attributes that explain most of the variability of
314 benzene exposures. A process known as CFS removed the redundant attributes in the data and allowed
315 for more interpretable data.

316

317 The models were trained on the RAS and were able to predict the classification of a participant to a PE
318 level based on just a few attributes in a similar fashion than using the FAS (as shown in Table S2). This
319 meant that CFS was able to remove the non-predictive attributes in the data. Thus only a few (most
320 predictive) attributes are needed to make an accurate prediction of the PE levels. Based on Table 2, the
321 predictive attributes common to all three PE campaigns could be grouped under the use of paints in

322 homes; upholstery materials; space heating and ETS. Although the paper focused on the results for
323 benzene as a VOC marker and as a known human carcinogen⁴⁴, the models are expected to give similar
324 results for the other VOCs, although some differences are seen⁶.

325

326 To assess the usefulness and practicality of the MLT models to predict and correctly classify PE to
327 benzene to be used in epidemiological studies, the performance of the models developed using MLTs
328 was analysed. For that purpose, different PE categories determined using percentiles, namely: High (>
329 90%ile), Medium (10-90%ile), and Low (< 10%ile); High (> 90%ile), Medium (30-90%ile), and Low
330 (< 30%ile), and High (> 70%ile), Medium (30-70%ile) and Low (< 30%ile) were compared.

331

332 MLTs were applied for the first time in PE modelling of benzene in comparison to linear regression
333 approaches, producing interesting results in the validation exercise where the test dataset was very small.
334 Nevertheless, further validation of the MLTs performance is required with larger datasets and for air
335 toxics that show different behaviour than benzene associated with their chemical composition, reactivity,
336 vapour pressure and indoor/outdoor dynamics. One earlier study⁴⁵ has predicted occupational exposure
337 to benzene in filling station workers using an ANN approach, and describing it as a promising technique.

338

339 All the MLT models used for this study proved to perform fairly well with better performance in the
340 Medium exposure ranges rather than in the Lower and Higher exposure ranges, whilst the GLM was
341 more predictive in the High exposure range. However, one should note that the low accuracies obtained
342 in the Low exposure range arose from the fact that the whole dataset was highly skewed to the lower
343 concentrations (Figure S2). Therefore, an even distribution of participants between all exposure level
344 classes would allow the models to estimate both the higher and lower exposure levels more accurately
345 as discussed hereunder.

346

347 Comparing the high exposure levels in MATCH, when the exposure category split is based on the 10-
348 90%iles or the 30-90%iles, (refer to Supporting Information, Table S7) ANN, NNGE and KStar perform
349 equally as the GLM in correctly classifying a maximum of 63% of the measured instances. On the other
350 hand, for EXPOLIS, the GLM fared much better than the abovementioned MLTs in correctly classifying
351 all high exposure instances. For TEACH in the 30-90%iles category ANN, NNGE and KStar were able
352 to classify only 33% of the measured instances whilst the GLM predicted all the measured instances.
353 However, when considering all the exposure categories and the number of cases correctly and incorrectly
354 classified, the overall performance of the models was very poor (Table 4), according to the proposed
355 rankings, making a large number of errors, which are penalised by the ranking proposed. Table 4 further
356 indicates that for appreciably large datasets, such as EXPOLIS, for both Epidemiology and Risk
357 Assessment applications, the MLTs ranked better with DT and RF appearing to be preferred in that order,
358 except when challenged with the 30-70%ile dataset. For smaller datasets, such as TEACH, the GLM
359 performed better, independently of the percentile classification used. However, when a 30-90%iles or
360 30-70%iles classification was used, the accuracy of all models (MLTs and GLM) in correctly classifying
361 cases decreased (Table 4) making a large number of classification errors.

362

363 The main goal of the regression model is to predict the assigned class (L, M or H) from the corresponding
364 attributes. It is important to stress the fact that when the Low category classification was changed from
365 the 10%ile to 30%ile, the number of samples in each category changed. In particular, this implied a larger
366 number of samples in the L bin. Since 75% and 25% of the samples from the entire dataset were randomly
367 selected for the training and testing of the models, the probability of picking a data point from the L class
368 increased, the probability of selecting a M sample decreased, while the probability of picking instances
369 from the H bin remained constant.

370

371 The performance of the MLTs is dependent on how training instances are distributed into the three
372 exposure categories and how the samples are randomly selected. Since sample selection is carried out
373 before each test run, the number of samples in each category (and hence the results shown in the
374 confusion matrices) can be different. Hence, in machine learning we cannot presume that the
375 performance on the H bin will remain the same (Table S6).

376

377 Two of the MLTs considered, namely KStar and ANN were also able to predict continuous data, as the
378 GLM does. From Table 5 it could be noted that interpreting the performance of the models, solely by
379 comparing R^2 can give an erroneous picture of the behaviour of the models. In this study, when
380 predicting continuous data, GLM performed better than MLTs. However, it can be concluded that for
381 cases where the dataset contains some missing values (such as in EXPOLIS), the KStar was found to be
382 an appreciably acceptable technique whereas for the cases where the dataset is quite small (such as
383 TEACH), the ANN seemed to have a comparable performance of a GLM. It was noted that GLM does
384 not seem to perform well for data which have very high or very low variance (such as tested for toluene
385 and 1,3-butadiene respectively but not discussed in this paper); an issue that is not crucial for the
386 robustness of the MLTs.

387

388 For the first time to our knowledge MLTs have been used to predict the PE of a person to air toxics such
389 as VOCs, in particular benzene, in this study. They appear to perform at least as well as the frequently
390 used GLM method and have the advantage of not requiring microenvironment concentration
391 measurements. In our earlier paper⁶, the dominant source of exposure to VOC including benzene were
392 road traffic, solvent use and ETS. This study identified important influences as use of paints in homes,
393 upholstery materials, space heating and ETS, and hence activity/lifestyles questionnaires should focus
394 on these sources additionally. The relative importance of each of these sources is likely to have changed

395 since the exposure studies used in this research were conducted, but they are still likely to influence
396 exposure heavily.

397

398 **ACKNOWLEDGEMENTS**

399 The authors thank all the 100 subjects who participated in MATCH. The data used in this research and
400 described in this article were obtained under contract to the Health Effects Institute (HEI), an organization
401 jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award R-
402 82811201) and certain motor vehicle and engine manufacturers. The contents of this article do not
403 necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies
404 of the EPA or motor vehicle and engine manufacturers. Special thanks go to Adam Gauci, Ian Fenech
405 Conti and Imran Sheikh for their fruitful discussion in preparing this paper.

406

407 **Conflict of Interests.** The authors declare no competing financial interest.

408

409 **Supporting Information.** Supporting Information provides further details of the Machine Learning
410 Techniques, the Correlation-Based Feature Subset, information on the distribution of personal
411 exposures, and detailed performance statistics and a Confusion Matrix for the models.

412

413 **REFERENCES**

414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460

1. Nieuwenhuijsen, M. J. (Ed.). Exposure assessment in occupational and environmental epidemiology, Oxford, UK. Oxford University Press 2003.
2. Nuckols, J. R.; Ward, M.; Jarup, L. Using geographic information systems for exposure assessment in environmental epidemiological studies. *Environ. Health Perspect.* 2004, 112, 1007-1015.
3. Nieuwenhuijsen, M. J.; Paustenbach, D.; Duarte-Davidson, R. New developments in exposure assessment: The impact on the practice of health risk assessment and epidemiological studies. *Environ. Int.* 2006, 32, 996-1009.
4. Burmaster, D. E.; Harris, R. H. The magnitude of compounding conservatisms in Superfund risk assessments. *Risk Anal.* 1993, 13, 131-134.
5. Nichols, A. L.; Zeckhauser, R. J. The perils of prudence: how conservative risk assessments distort regulation. *Regulat. Toxicol. Pharmacol.* 1988, 8, 61-75.
6. Delgado-Saborit, J. M.; Aquilina, N. J.; Meddings, C.; Baker, S.; Harrison, R. M. Model development and validation of personal exposure to volatile organic compound concentrations. *Environ. Health Perspect.* 2009b, 117, 1571-1579.
7. Heavner, D. L.; Morgan, W. T.; Ogden, M. W. Determination of volatile organic-compounds and ETS apportionment in 49 homes. *Environ. Int.* 1995, 21, 3-21.
8. Austin, C. C.; Wang, D.; Ecobichon, D. J.; Dussault, G. Characterization of volatile organic compounds in smoke at experimental fires. *J. Toxicol. Environ. Health* 2001, 63, 191-206.
9. Ilgen, E.; Karfich, N.; Levsen, K.; Angerer, J.; Schneider, P.; Heinrich, J.; Wichmann, H.-E.; Dunemann, L.; Begerow, J. (). Aromatic hydrocarbons in the atmospheric environment: part I. Indoor versus outdoor sources, the influence of traffic. *Atmos. Environ.* 2001, 35, 1235-1252.
10. Yang, P. Y.; Lin, J. L.; Hall, A. H.; Tsao, T. C. Y.; Chern, M. S. Acute ingestion poisoning with insecticide formulations containing the pyrethroid permethrin, xylene, and surfactant: a review of 48 cases. *J. Toxicol. Clin. Toxicol.* 2002, 40, 107-113.
11. Edwards, R. D.; Schweizer, C.; Jantunen, M.; Lai, H. K.; Bayer-Oglesby, L.; Katsouyanni, K. Personal exposures to VOC in the upper end of the distribution – relationships to indoor, outdoor and workplace concentrations. *Atmos. Environ.* 2005, 39, 2299-2307.
12. Batterman, S.; Jia, C. R.; Hatzivasilis, G.; Godwin, C. Simultaneous measurement of ventilation using tracer gas techniques and VOC concentrations in homes, garages and vehicles. *J. Environ. Monitor.* 2006, 8, 249-256.
13. Curren, K. C.; Dann, T. F.; Wang, D. K. Ambient air 1,3-butadiene concentrations in Canada (1995–2003): seasonal, day of week variations, trends, and source influences. *Atmos. Environ.* 2006, 40, 170-181.

- 461 14. Zuraimi, M. S.; Roulet, C. A.; Tham, K. W.; Sekhar, S. C.; Cheong, K. W. D.; Wong, N. H.; Lee,
462 K. H. A comparative study of VOCs in Singapore and European office buildings. *Build. Environ.*
463 2006, 41, 316-329.
- 464
- 465 15. Song, Y.; Shao, M.; Liu, Y.; Lu, S.H.; Kuster, W.; Goldan, P.; Shaodong, X. Source apportionment
466 of ambient volatile organic compounds in Beijing. *Environ. Sci. Technol.* 2007, 41, 4348-4353.
- 467
- 468 16. Ordieres, J. B.; Vergara, E. P.; Capuz, R. S.; Salazar, R. E. Neural network prediction model for fine
469 particulate matter (PM_{2.5}) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez
470 (Chihuahua). *Environ. Model. Softw.* 2005, 20, 547-559.
- 471
- 472 17. Perez, P.; Reyes, J. Prediction of maximum of 24-h average of PM₁₀ concentrations 30 h in advance
473 in Santiago, Chile. *Atmos. Environ.* 2002, 36, 4555-4561.
- 474
- 475 18. M.; Niska, H.; Dorling, S.; Chatterton, T.; Foxall, R.; Cawley, G. Extensive evaluation of neural
476 network models for the prediction of NO₂ and PM₁₀ concentrations, compared with a deterministic
477 modelling system and measurements in central Helsinki. *Atmos. Environ.* 2003, 37, 4539-4550.
- 478
- 479 19. Lu, W. Z.; Wang, W. J.; Wang, X. K.; Xu, Z. B.; Leung, A. Y. T. Using improved neural network
480 model to analyse RSP, NO_x and NO₂ levels in urban air in Mong Kok, Hong Kong. *Environ. Monitor.*
481 & Assess. 2003, 87, 235-254.
- 482
- 483 20. Hooyberghs, J.; Mensink, C.; Dumont, G.; Fierens, F.; Brasseur, O. A neural network forecast for
484 daily average PM₁₀ concentrations in Belgium. *Atmos. Environ.* 2005, 39, 3279-3289.
- 485
- 486 21. Raimondo, G.; Montuori, A.; Moniaci, W.; Pasero, E.; Almkvist, E. An application of machine
487 learning methods to PM₁₀ levels medium-term prediction. Proceedings of the 11th international
488 conference, KES 2007 and XVII Italian workshop on neural networks conference on Knowledge-
489 based intelligent information and engineering systems 2007b, Part III, 259-266.
- 490
- 491 22. Paschalidou, A. K.; Karakitsios, S.; Kleanthous, S.; Kassomenos, P. A. Forecasting hourly PM₁₀
492 concentration in Cyprus through artificial neural networks and multiple regression models:
493 implications to local environment management. *Environ. Sci. Polluti. Res.* 2010, 18, 316-327.
- 494
- 495 23. Raimondo, G.; Montuori, A.; Moniaci, W.; Pasero, E.; Almkvist, E. A Machine Learning Tool to
496 Forecast PM₁₀ Level. 5th Conference on Artificial Intelligence Applications to Environmental
497 Science. AMS 87th Annual Meeting 2007a, 14-18 January 2007, San Antonio, TX.
- 498
- 499 24. Aquilina, N. J.; Delgado Saborit, J. M.; Gauci, A. P.; Baker, S.; Meddings, C.; Harrison, R. M.
500 Comparative modeling approaches for personal exposure to particle-associated PAH. *Environ.Sci.*
501 *Technol.* 2010, 44, 9370-9376.
- 502
- 503 25. Sarigiannis, D. A.; Karakitsios, S. P.; Gotti, A.; Papaloukas, C. L.; Kassomenos, P.A.; Pilidis, G. A.
504 Bayesian Algorithm Implementation in a Real Time Exposure Assessment Model on Benzene with
505 Calculation of Associated Cancer Risks; *Sensors* 2009, 9, 731-755.
- 506
- 507 26. Delgado-Saborit, J. M.; Aquilina, N. J; Meddings, C.; Baker, S.; Vardoulakis, S.; Harrison, R. M.
508 Measurement of personal exposure to volatile organic compounds and particle associated PAH in
509 three UK regions. *Environ. Sci. Technol.* 2009a, 43, 4582-4588.

- 510
511 27. Hanninen, O.; Alm, S.; Kaarakainen, E.; Jantunen, M. The EXPOLIS databases. Kuopio 2002.
512
513 28. Kinney, P. L.; Chillrud, S. N.; Sonja, R.; James, R.; Spengler, J. D. Exposures to Multiple Air Toxics
514 in New York City; *Environ. Health Perspect.* 2002, 110, 539-546.
515
516 29. Harrison, R. M.; Delgado Saborit, J. M.; Baker, S. J.; Aquilina, N.; Meddings, C.; Harrad, S.;
517 Matthews, I.; Vardoulakis, S.; Anderson, H. R. Measurement and modeling of exposure to selected
518 air toxics for health effects studies and verification by biomarkers. HEI Research Report 143, Health
519 Effects Institute 2009, Boston, MA.
520
521 30. Kinney, P. L.; Chillrud, S. N.; Sax, S.; Ross, J. M.; Pederson, D. C.; Johnson, D.; Aggarwal, M;
522 Spengler, J. D. Toxic Exposure Assessment: A Columbia-Harvard (TEACH) Study (The New York
523 City Report). NUATRC research report, no. 3. 2005, Air Toxics Research Centre. Available online:
524 http://teach.aer.com/docs/NY_TEACH_Study3.pdf
525
526 31. Hall, M. A.; Smith, L. A. Practical feature subset selection for machine learning. McDonald, C.
527 (Ed.), *Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference*
528 *ACSC'98*, Perth, Australia, 4-6 February, Springer 1998, 181-191.
529
530 32. Chang, M.-H.; Ha, E.-H.; Park, H.; Ha, M.; Kim, Y. J.; Hong, Y.-C.; Kim, Y.; Roh, Y.-M.; Lee, B.-
531 E.; Seo, J.-H; Kim, B.-M. The Effect of VOCs Exposure During Pregnancy on Newborn's Birth
532 Weight in Mothers and Children's Environmental Health (MOCEH) Study. *Epidemiol.* 2011, 22,
533 S162-S163.
534
535 33. Delgado-Saborit, J. M.; Aquilina, N.; Meddings, C.; Bakar, S.; Harrison, R.M. Relationship of
536 personal exposure to volatile organic compounds to home, work and fixed site outdoor
537 concentrations. *Sci. Tot. Environ.* 2011, 409, 478-488.
538
539 34. Lee, J.-W.; Jo, W.-K. Actual commuter exposure to methyl-tertiary butyl ether, benzene and toluene
540 while traveling in Korean urban areas. *Sci.Tot. Environ.* 2002, 291, 219-228.
541
542 35. Violante, F. S.; Sanguinetti, G.; Barbieri, A.; Accorsi, A.; Mattioli, S.; Cesari, R.; Fimognari, C.;
543 Hrelia, P. Lack of correlation between environmental or biological indicators of benzene exposure
544 at parts per billion levels and micronuclei induction. *Environ. Res.* 2003, 91, 135-142.
545
546 36. Carrington, J.; Watson, A. F. R.; Gee, I. L. The effects of smoking status and ventilation on
547 environmental tobacco smoke concentrations in public areas of UK pubs and bars. *Atmos. Environ.*
548 2003, 37, 3255-3266.
549
550 37. Ozcift, A.; Gulden, A. Classifier Ensemble Construction with Rotation Forest to Improve Medical
551 Diagnosis Performance of Machine Learning Algorithms. *Comput. Methods Programs Biomed.*
552 2011, 104 (3), 443-451.
553
554 38. Li, M.; Zhou, Z. H. Improve Computer-Aided Diagnosis with Machine Learning Techniques Using
555 Undiagnosed Samples. *IEEE Trans. Syst. Man, Cybern. Part A Systems Humans* 2007, 37 (6), 1088-
556 1098.
557

- 558 39. Lee, M. C.; Boroczky, L.; Sungur-Stasik, K.; Cann, A. D.; Borczuk, A. C.; Kawut, S. M.; Powell,
559 C. A. A Two-Step Approach for Feature Selection and Classifier Ensemble Construction in
560 Computer-Aided Diagnosis. In *Proceedings - IEEE Symposium on Computer-Based Medical*
561 *Systems*; 2008; pp 548–553.
- 562
- 563 40. WEKA. Description of Decision Trees from WEKA website 2010, <http://wekadocs.com/node/2>.
564 Accessed on 07/09/2010.
- 565
- 566 41. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data
567 Mining Software: An Update. *SIGKDD Explorations* 2009, 11.
- 568
- 569 42. Wallace, L. A. Major sources of benzene exposure. *Environ. Health Perspect.* 1989, 82, 165-169.
- 570
- 571 43. Batterman, S.; Godwin, C.; C. Jia. Long duration tests of room air filters in cigarette smokers'
572 homes. *Environ. Sci. Technol.*, 2005 39, 7260-7268.
- 573
- 574 44. IARC. Agents classified by the IARC monographs, 2011, Volumes 1-100. Available online:
575 <http://monographs.iarc.fr/ENG/Classification/ClassificationsAlphaOrder.pdf>.
- 576
- 577 45. Karakitsios, S. P.; Papaloukas, C. L.; Kassomenos, P. A.; Pilidis, G. A. Assessment and prediction
578 of exposure to benzene of filling station employees. *Atmos. Environ.*, 41, 9555-9569.
- 579
- 580
- 581
- 582
- 583
- 584
- 585

586 **TABLE LEGENDS**

587

588 **Table 1:** The bin limit values for benzene (in $\mu\text{g m}^{-3}$) determined by the 10%ile and 90%ile, 30%ile
589 and 70%ile and the 30%ile and 90%ile percentiles.

590

591 **Table 2:** Reduced number of attributes (RAS) using the CFS algorithm, which are able to predict the
592 continuous benzene concentration for (a) MATCH, (b) EXPOLIS, (c) TEACH.

593

594 **Table 3:** Point ranking system devised for our models if they are to be used in epidemiology and risk
595 assessment applications to predict benzene correctly in three studies.

596

597 **Table 4:** Ranking of the different models' performance to predict benzene correctly in three studies.
598 Numbers in bold indicate the models which ranked highest in correctly classifying
599 instances in L, M and H exposure categories.

600

601 **Table 5:** Predicting continuous data results for benzene.

602

603

604