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

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

Noel J. Aquilina[1,2], Juana Maria Delgado-Saborit[1‡], Stefano Bugelli[3], Jason Padovani Ginies[3] and Roy M. Harrison[1*†],

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

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

[3]Department 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

ABSTRACT

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

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

1. INTRODUCTION

Exposure assessment is an important analytical tool for evaluating the likelihood and extent of actual or potential exposure of people to pollutants and is an important component of any health risk assessment and epidemiological study. Exposure to chemicals from environmental and occupational settings can be characterized in different ways[1]. Direct methods such as personal monitoring and biomarkers are considered to be accurate for exposure assessment yet are costly to study big populations. Indirect information gained through questionnaires and diaries accompanied by environmental monitoring can be used to develop exposure models. Modelling techniques have greatly improved the assessments and are likely to be important in future studies since direct measurement of exposure is often too expensive and time consuming.

In recent years, exposure assessment to atmospheric pollutants has been conducted mainly either by deterministic methods, strengthened by geographical information systems and geostatistical techniques[2], or by a statistical approach[3]. In the last 20 years statistical approaches have focused on regression techniques and source apportionment while probabilistic modelling was mainly done by Monte Carlo analyses and Bayesian statistics. The main criticisms of many exposure assessments have been a reliance on overly conservative assumptions about exposure, as well as the problem of how to model properly the highly exposed populations that generally are small in number[4,5]. The earlier published work has shown a limited ability of methods based upon measurement of microenvironment concentrations to provide an accurate quantitative reconstruction of personal exposure (PE). This is no doubt due to the variability in concentrations within a given type of microenvironment and poorly quantified contributions from sporadic sources. Since machine learning techniques (MLTs) function without *a priori* assumptions of pathways and have great power to extract meaningful patterns and trends from datasets, we have for the first time applied MLTs to the modelling of non-occupational PE to a key air pollutant, benzene.

Ideally a PE model should be able to predict the degree of exposure of an individual based on a minimum number of input attributes. The model for benzene developed by Delgado-Saborit et al.[6] predicted the PE by integrating the time fraction spent in each microenvironment times the concentration of benzene in the microenvironment visited, and also accounted for external factors that might affect exposure as add-on variables, using a linear regression approach. The best model that was able to predict PE with independence of measurements was based upon certain time-activity attributes. Other studies conducted 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 et al.[13], Zuraimi et al.[14] and Song et al.[15], through source apportionment, have identified sources of benzene that were consistent with the variables that were introduced in the above-mentioned model. The model identified the most important non-weather-related variables for benzene exposures, highlighting the influence of personal activities, use of solvents, and exposure to environmental tobacco smoke (ETS) on PE levels.

MLTs are used for several air quality applications, including forecasting of airborne pollutants such as $PM_{2.5}$ levels[16], PM_{10} levels[17,18,19,20,21,22], SO_2, CO and NO and NO_2 and O_3[19,23], and particle-phase PAH[24]. One study uses a MLT to model benzene exposures, but in an occupational setting[25].

In this study, MLT models were trained and tested on benzene PE data that was collected during three PE campaigns, namely; MATCH[26], TEACH[27] and EXPOLIS[28]. The performance of the MLT models in classifying personal exposures was tested and results are discussed in the light of their usefulness for risk assessment and epidemiological studies.

2. METHODOLOGY

2.1 Description of Datasets

Three datasets were employed in training and testing the models using MLTs. These datasets as described in detail below were the MATCH, the EXPOLIS and the TEACH databases. Descriptive statistics appear in Table S1 and Figure S3.

The MATCH (**M**easurement and Modelling of **A**ir **T**oxics **C**oncentrations for **H**ealth Studies) study's main objective was to optimize a model of PE based on microenvironment concentrations and time/activity diaries and to compare the modelled with measured exposures in an independent dataset[6]. The subjects for this study, enrolled to measure their PE to a suite of air toxics were recruited based upon a set of inclusion determinants that affected exposure, namely: location, living in houses with heavy trafficked roads (termed as first line houses), having a house with an integral garage, and exposure to ETS[26]. PE of 100 adult non-smokers living in three UK locations, namely London, West Midlands, and rural South Wales, to 15 VOCs was measured using an actively pumped sampler carried around by the subjects for five consecutive 24 hr periods, following their normal lifestyle.

The EXPOLIS (Air Pollution Exposure Distributions within Adult Urban Populations in Europe) study focused on adults living in cities in seven European countries (Helsinki, Athens, Basel, Grenoble, Milan, Prague, Oxford), exposed to air pollutants in their homes, workplaces and other common urban micro-environments[27] from 1996-1998. The 401 subjects who participated in this study were chosen according to certain criteria which are found in the EXPOLIS manual[27]. This study was based on a single 48 hr sampling period using a suitcase containing the sampler.

The TEACH (**T**oxic **E**xposure **A**ssessment, a **C**olumbia / **H**arvard) study was designed to characterize levels and factors of PE to urban air toxics among high school students in Los Angeles and New York from 1999-2000[28]. This study involved 87 students who carried a backpack for 48 hr over two different sampling periods, one in summer and another in winter.

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

2.2 Attribute Selection for dimension reduction

Attribute subset selectors are a collection of algorithms that try to find and remove irrelevant and redundant attributes[31], an exercise termed as dimension reduction that is required in generating robust PE models requiring a minimal number of attributes.

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

3. GENERAL LINEAR MODELLING TO MODEL PE TO BENZENE

A more common approach to modelling PE is by using a General Linear Model (GLM) which was used in various studies, such as to model the effect of VOCs exposure during pregnancy to newborn's birth weight[32], to find the relationship between PE to VOCs and home, work and outdoor concentrations[33], to evaluate vehicle exposure to certain VOCs including benzene in urban areas[34]. In this study a GLM was developed and compared with the MLTs described in Section 4.

The GLM is a combination of two major model types, namely regression models and analysis of variance models. For this study, where only one dependent (continuous) variable was available, GLMs were used. Here, all the attributes were included into the model and the least significant was removed manually one at a time. This process was repeated until the remaining variables left were all statistically significant ($p<0.05$). This was also used in previous exposure studies such as benzene exposure[35] and exposure to ETS[36].

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

4. MACHINE LEARNING TECHNIQUES TO MODEL PE TO BENZENE

Our earlier research[6] was based upon the use of simple additive models in which microenvironment concentrations were summed in a time-weighted manner, or multiple linear regression approaches in which key influences upon exposure were identified and added in weighted manner to obtain the best overall fit to the measured exposures. Such methods require *a priori* assumptions as to the most important factors/sources influencing exposure and assume that total exposure is the linear sum of a range of weighted contributions.

MLTs used in this study are computer-based algorithms which recognise features in datasets which when combined give a good fit to an outcome variable, in this case the measured PE. The algorithms learn directly from the data and improve their performance as they are provided with more samples. MLTs can be either supervised or unsupervised. In the former case, a known set of input data and output responses is used to combine input variables in such a way as to predict the outcome using classification or regression methods. In the unsupervised learning case, methods such as clustering are used to recognise patterns in the data without reference to the outputs.

In several applications predictions have been aided by the application of MLTs[37]. Algorithms are generally trained with previously available data and allow predictions in the testing phase[38]. The success of an analysis can thus be defined as the ability of such algorithms to predict the correct status of unseen data.

In the realm of PE to atmospheric pollutants, accuracy of classification strategies can be affected negatively with the use of too many features in the classification. This may lead to overfitting, in which noise or irrelevant features may decrease classification accuracy because of the finite size of the training samples[39]. The mining workbench program used for developing the MLT models was the Waikato Environment for Knowledge Analysis (WEKA)[40,41]. Further information on the MLTs used in this research is given in the Supplementary Information.

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

5. MODELS AND CLASSIFICATION OF EXPOSURE

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

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

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

Table 1: The bin limit values for benzene (in µg m^{-3}) determined by the 10%ile and 90%ile, 30%ile and 70%ile and the 30%ile and 90%ile percentiles.

	Low (L)		**Medium (M)**			**High (H)**	
Study	**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

6. RESULTS

6.1 Testing Attribute Selection and Accuracy of Classification

ACFS algorithm was used to remove irrelevant and redundant variables from a Full Attribute Set (FAS). A RAS for each study was obtained and the important attributes identified by CFS were compared with similar attributes identified in other studies and are summarized in Table 2.

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

(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]

(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	

(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	

In order to assess the performance of the MLTs, these were run using the FAS and the RAS from the three studies, where the RAS was obtained by CFS as explained above. Table S3 summarizes the overall accuracy obtained for predicting PE to benzene when using the FAS and the RAS for classification.

The overall performance of the MLTs in a 10-fold cross validation and a 25% testing dataset using a 75% training dataset for classification determined by 10 and 90 percentiles using the RAS are presented in Tables S4 and S5 respectively. The accuracy for the MLTs was calculated via a confusion matrix available in WEKA that was generated in order to compare the various models used in trying to predict PE (Supporting Information, Table S6). The matrix, for each model used, summarizes the correctly classified instances and also indicates in which category the model wrongly classified instances when compared to the corresponding measured instances. The degree of accuracy of the models can then be determined by calculating the percentage of instances correctly classified and attributed to the correct concentration range bin. Table S7 compares the performance of the MLTs with the GLM in correctly classifying the exposure classes.

If these models are to be used for epidemiology or risk assessment applications, the need for correct classification of the PE in different exposure categories varies according to the choice of the percentile ranges chosen in this paper (10-90, 30-70 and 30-90%iles). A point ranking system (Table 3) has been devised for the abovementioned applications and applied to the confusion matrix (Table S6) in order to identify which model scores best in classifying the modelled concentrations in the correct classification categories (L, M and H) as the corresponding measured concentrations. Table 4 shows the total ranking of each model based on the point ranking system summarised in Table 3.

The scoring scheme for epidemiology applications penalised extreme misclassification highly (i.e. H to L and L to H), and lesser misclassification less harshly with incorrect prediction of M as L or H losing more points than the reverse error. The rationale was that epidemiology depends heavily upon a gradient of exposures in which the H and L are most important in defining the distribution.

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

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

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**

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

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

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

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

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

While the majority of the MLTs predict only exposure category, two of the MLTs (KStar and ANN) and the GLM were able to predict also continuous data. The R^2 value and the Predicted vs Measured gradient are shown in Table 5. DT, NNGE and RF are not included as they do not give R^2 values for direct comparison with the GLM. This table indicates that the performance of the model is not determined solely by the R^2 value; in fact, the predicted: measured ratio indicates that the GLM perform better in predicting a PE value closer to the measured values when compared to the MLTs, at least in the studies considered.

Table 5: Predicting continuous data results for benzene.

Study	Model	Predicted : Measured Ratio	R^2
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

7. DISCUSSION

This study presents several PE models developed using different MLTs using benzene PE data collected during three independent PE campaigns, namely; MATCH[26], and EXPOLIS[27] and TEACH[28]. The first step in the model development was to select those attributes that explain most of the variability of benzene exposures. A process known as CFS removed the redundant attributes in the data and allowed for more interpretable data.

The models were trained on the RAS and were able to predict the classification of a participant to a PE level based on just a few attributes in a similar fashion than using the FAS (as shown in Table S2). This meant that CFS was able to remove the non-predictive attributes in the data. Thus only a few (most predictive) attributes are needed to make an accurate prediction of the PE levels. Based on Table 2, the predictive attributes common to all three PE campaigns could be grouped under the use of paints in homes; upholstery materials; space heating and ETS. Although the paper focused on the results for benzene as a VOC marker and as a known human carcinogen[44], the models are expected to give similar results for the other VOCs, although some differences are seen[6].

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

MLTs were applied for the first time in PE modelling of benzene in comparison to linear regression approaches, producing interesting results in the validation exercise where the test dataset was very small. Nevertheless, further validation of the MLTs performance is required with larger datasets and for air toxics that show different behaviour than benzene associated with their chemical composition, reactivity, vapour pressure and indoor/outdoor dynamics. One earlier study[45] has predicted occupational exposure to benzene in filling station workers using an ANN approach, and describing it as a promising technique.

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

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

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

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

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

For the first time to our knowledge MLTs have been used to predict the PE of a person to air toxics such as VOCs, in particular benzene, in this study. They appear to perform at least as well as the frequently used GLM method and have the advantage of not requiring microenvironment concentration measurements. In our earlier paper[6], the dominant source of exposure to VOC including benzene were road traffic, solvent use and ETS. This study identified important influences as use of paints in homes, upholstery materials, space heating and ETS, and hence activity/lifestyles questionnaires should focus on these sources additionally. The relative importance of each of these sources is likely to have changed since the exposure studies used in this research were conducted, but they are still likely to influence exposure heavily.

ACKNOWLEDGEMENTS

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

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

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

REFERENCES

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.

14. Zuraimi, M. S.; Roulet, C. A.; Tham, K. W.; Sekhar, S. C.; Cheong, K. W. D.; Wong, N. H.; Lee, K. H. A comparative study of VOCs in Singapore and European office buildings. Build. Environ. 2006, 41, 316-329.

15. Song, Y.; Shao, M.; Liu, Y.; Lu, S.H.; Kuster, W.; Goldan, P.; Shaodong, X. Source apportionment of ambient volatile organic compounds in Beijing. Environ. Sci. Technol. 2007, 41, 4348-4353.

16. Ordieres, J. B.; Vergara, E. P.; Capuz, R. S.; Salazar, R. E. Neural network prediction model for fine particulate matter ($PM_{2.5}$) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environ. Model. Softw. 2005, 20, 547-559.

17. Perez, P.; Reyes, J. Prediction of maximum of 24-h average of PM_{10} concentrations 30 h in advance in Santiago, Chile. Atmos. Environ. 2002, 36, 4555-4561.

18. M.; Niska, H.; Dorling, S.; Chatterton, T.; Foxall, R.; Cawley, G. Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmos. Environ. 2003, 37, 4539-4550.

19. Lu, W. Z.; Wang, W. J.; Wang, X. K.; Xu, Z. B.; Leung; A. Y. T. Using improved neural network model to analyse RSP, NO_x and NO_2 levels in urban air in Mong Kok, Hong Kong. Environ. Monitor. & Assess. 2003, 87, 235-254.

20. Hooyberghs, J.; Mensink, C.; Dumont, G.; Fierens, F.; Brasseur, O. A neural network forecast for daily average PM10 concentrations in Belgium. Atmos. Environ. 2005, 39, 3279-3289.

21. Raimondo, G.; Montuori, A.; Moniaci, W.; Pasero, E.; Almkvist, E. An application of machine learning methods to PM10 levels medium-term prediction. Proceedings of the 11th international conference, KES 2007 and XVII Italian workshop on neural networks conference on Knowledge-based intelligent information and engineering systems 2007b, Part III, 259-266.

22. Paschalidou, A. K.; Karakitsios, S.; Kleanthous, S.; Kassomenos, P. A. Forecasting hourly PM10 concentration in Cyprus through artifical neural networks and multiple regression models: implications to local environment management. Environ. Sci. Polluti. Res. 2010, 18, 316-327.

23. Raimondo, G.; Montuori, A.; Moniaci, W.; Pasero, E.; Almkvist, E. A Machine Learning Tool to Forecast PM_{10} Level. 5th Conference on Artificial Intelligence Applications to Environmental Science. AMS 87th Annual Meeting 2007a, 14-18 January 2007, San Antonio, TX.

24. Aquilina, N. J.; Delgado Saborit, J. M.; Gauci, A. P.; Baker, S.; Meddings, C.; Harrison, R. M. Comparative modeling approaches for personal exposure to particle-associated PAH. Environ.Sci. Technol. 2010, 44, 9370-9376.

25. Sarigiannis, D. A.; Karakitsios, S. P.; Gotti, A.; Papaloukas, C. L.; Kassomenos, P.A.; Pilidis, G. A. Bayesian Algorithm Implementation in a Real Time Exposure Assessment Model on Benzene with Calculation of Associated Cancer Risks; Sensors 2009, 9, 731-755.

26. Delgado-Saborit, J. M.; Aquilina, N. J; Meddings, C.; Baker, S.; Vardoulakis, S.; Harrison, R. M. Measurement of personal exposure to volatile organic compounds and particle associated PAH in three UK regions. Environ. Sci. Technol. 2009a, 43, 4582-4588.

27. Hanninen, O.; Alm, S.; Kaarakainen, E.; Jantunen, M. The EXPOLIS databases. Kuopio 2002.

28. Kinney, P. L.; Chillrud, S. N.; Sonja, R.; James, R.; Spengler, J. D. Exposures to Multiple Air Toxics in New York City; Environ. Health Perspect. 2002, 110, 539-546.

29. Harrison, R. M.; Delgado Saborit, J. M.; Baker, S. J.; Aquilina, N.; Meddings, C.; Harrad, S.; Matthews, I.; Vardoulakis, S.; Anderson, H. R. Measurement and modeling of exposure to selected air toxics for health effects studies and verification by biomarkers. HEI Research Report 143, Health Effects Institute 2009, Boston, MA.

30. Kinney, P. L.; Chillrud, S. N.; Sax, S.; Ross, J. M.; Pederson, D. C.; Johnson, D.; Aggarwal, M; Spengler, J. D. Toxic Exposure Assessment: A Columbia-Harvard (TEACH) Study (The New York City Report). NUATRC research report, no. 3. 2005, Air Toxics Research Centre. Available online: http://teach.aer.com/docs/NY_TEACH_Study3.pdf

31. Hall, M. A.; Smith, L. A. Practical feature subset selection for machine learning. McDonald, C. (Ed.), Computer Science ’98 Proceedings of the 21st Australasian Computer Science Conference ACSC’98, Perth, Australia, 4-6 February, Springer 1998, 181-191.

32. Chang, M.-H.; Ha, E.-H.; Park, H.; Ha, M.; Kim, Y. J.; Hong, Y.-C.; Kim, Y.; Roh, Y.-M.; Lee, B.-E.; Seo, J.-H; Kim, B.-M. The Effect of VOCs Exposure During Pregnancy on Newborn's Birth Weight in Mothers and Children's Environmental Health (MOCEH) Study. Epidemiol. 2011, 22, S162-S163.

33. Delgado-Saborit, J. M.; Aquilina, N.; Meddings, C.; Bakar, S.; Harrison, R.M. Relationship of personal exposure to volatile organic compounds to home, work and fixed site outdoor concentrations. Sci. Tot. Environ. 2011, 409, 478–488.

34. Lee, J.-W.; Jo, W.-K. Actual commuter exposure to methyl-tertiary butyl ether, benzene and toluene while traveling in Korean urban areas. Sci.Tot. Environ. 2002, 291, 219-228.

35. Violante, F. S.; Sanguinetti, G.; Barbieri, A.; Accorsi, A.; Mattioli, S.; Cesari, R.; Fimognari, C.; Hrelia, P. Lack of correlation between environmental or biological indicators of benzene exposure at parts per billion levels and micronuclei induction. Environ. Res. 2003, 91, 135-142.

36. Carrington, J.; Watson, A. F. R.; Gee, I. L. The effects of smoking status and ventilation on environmental tobacco smoke concentrations in public areas of UK pubs and bars. Atmos. Environ. 2003, 37, 3255-3266.

37. Ozcift, A.; Gulten, A. Classifier Ensemble Construction with Rotation Forest to Improve Medical Diagnosis Performance of Machine Learning Algorithms. *Comput. Methods Programs Biomed.* 2011, *104* (3), 443–451.

38. Li, M.; Zhou, Z. H. Improve Computer-Aided Diagnosis with Machine Learning Techniques Using Undiagnosed Samples. *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans* **2007**, *37* (6), 1088–1098.

39. Lee, M. C.; Boroczky, L.; Sungur-Stasik, K.; Cann, A. D.; Borczuk, A. C.; Kawut, S. M.; Powell, C. A. A Two-Step Approach for Feature Selection and Classifier Ensemble Construction in Computer-Aided Diagnosis. In *Proceedings - IEEE Symposium on Computer-Based Medical Systems*; 2008; pp 548–553.

40. WEKA. Description of Decision Trees from WEKA website 2010, http://wekadocs.com/node/2. Accessed on 07/09/2010.

41. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009, 11.

42. Wallace, L. A. Major sources of benzene exposure. Environ. Health Perspect. 1989, 82, 165-169.

43. Batterman, S.; Godwin, C.; C. Jia. Long duration tests of room air filters in cigarette smokers' homes. Environ. Sci.Technol., 2005 39, 7260-7268.

44. IARC. Agents classified by the IARC monographs, 2011, Volumes 1-100. Available online: http://monographs.iarc.fr/ENG/Classification/ClassificationsAlphaOrder.pdf.

45. Karakitsios, S. P.; Papaloukas, C. L.; Kassomenos, P. A.; Pilidis, G. A. Assessment and prediction of exposure to benzene of filling station employees. Atmos. Environ., 41, 9555-9569.

TABLE LEGENDS

Table 1: The bin limit values for benzene (in μg m^{-3}) determined by the 10%ile and 90%ile, 30%ile and 70%ile and the 30%ile and 90%ile percentiles.

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

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

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.

Table 5: Predicting continuous data results for benzene.