Environ. Sci. Technol. **2010,** *44,* 9370–9376

Comparative Modeling Approaches for Personal Exposure to Particle-Associated PAH

NOEL J. AQUILINA,[†,‡] JUANA MARI DELGADO-SABORIT,[†] ADAM P. GAUCI,[§] STEPHEN BAKER,[†] CLAIRE MEDDINGS,[†] AND ROY M. HARRISON*[,†]

Division of Environmental Health and Risk Management, School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham B15 2TT, United Kingdom, Department of Physics, Faculty of Science, University of Malta, Msida MSD 2080, Malta, and Department of Intelligent Computer Systems, Faculty of ICT, University of Malta, Msida MSD 2080, Malta

Received July 26, 2010. Revised manuscript received October 28, 2010. Accepted November 1, 2010.

Several models for simulation of personal exposure (PE) to particle-associated polycyclic aromatic hydrocarbons (PAH) have been developed and tested. The modeling approaches include linear regression models (Model 1), time activity weighted models (Models 2 and 3), a hybrid model (Model 4), a univariate linear model (Model 5), and machine learning technique models (Model 6 and 7). The hybrid model (Model 4), which utilizes microenvironment data derived from time-activity diaries (TAD) with the implementation of add-on variables to account for external factors that might affect PE, proved to be the best regression model (R^2 for B(a)P = 0.346, $p < 0.01$; N = 68). This model was compared with results from two machine learning techniques, namely decision trees (Model 6) and neural networks (Model 7), which represent an innovative approach to PE modeling. The neural network model was promising in giving higher correlation coefficient results for all PAH (R^2 for B(a)P = 0.567, $p < 0.01$; N = 68) and good performance with the smaller test data set (R^2 for B(a)P = 0.640, $p < 0.01$; N = 23). Decision tree accuracies (Model 6) which assess how precisely the algorithm can determine the correct classification of a PE concentration range indicate good performance, but this is not comparable to the other models through R^2 values. Using neural networks (Model 7) showed significant improvements over the performance of hybrid Model 4 and the univariate general linear Model 5 for test samples (not used in developing the models). The worst performance was given by linear regression Models 1 to 3 based solely on home and workplace concentrations and time-activity data.

1. Introduction

Airborne particle-associated polycyclic aromatic hydrocarbons (PAH) are produced by high-temperature reactions such as incomplete combustion and pyrolysis of fossil fuels and other organic materials (*1, 2*). Nonoccupational sources of PAH exposure can be subdivided between outdoor and indoor. The outdoor sources include exhaust from motor vehicles (*3*), tire wear debris, asphalt particles (*4*), forest fires, and coal and oil burning (*3*). In indoor environments, PAH are generated in fumes from various cooking oils (*5*), meat cooking (*6, 7*), domestic wood burning (*8*), and from coal briquettes and charcoal for domestic heating (*9*). Natural gas home appliances, candles, incense burning (*10*), and glue containing coal-tar pitch in the underlay of parquet floors (*11*) were identified as other indoor sources. PAH are also transferred inside from outdoors either from nearby sources or due to long-range transport (*12, 13*). Nevertheless, an important contribution to PAH indoor air levels comes from environmental tobacco smoke (ETS) (*7, 14*).

Delgado-Saborit et al. (*15*) reported that PE to PAH may be estimated from *a*) centrally located monitors, *b*) from the combination of fixed-point monitors in individual microenvironments and activity data defining the times spent in each of the microenvironments, or *c*) from direct PE monitoring (*16, 17*). Liu et al. (*18*) argue that direct measurement of PE to PAH via personal monitoring is the most accurate exposure assessment method currently available; however, carrying out personal monitoring at the population level is costly and impractical. Serrano-Trespalacios et al. (*19*) and Wilson et al. (*20*) report that centrally located monitors have a tendency to underestimate exposures. Consequently, indirect estimation via a combination of microenvironment concentrations and personal time/activity diaries is an attractive alternative. Earlier studies showed that modeled PE can provide a good prediction of overall measured PE (*21*), and this offers an effective means of estimating PE to these pollutants without the considerable logistical difficulties of personal sampling.

Previous studies have quantified the level of contribution of each microenvironment or activity to the total PE in the nonoccupationally exposed population and have identified sources affecting PE (*22*−*24*). The locations where time is spent; the activity patterns, which determine the time spent in different microenvironments; the type of activities in which people are involved (*21, 25*); sociodemographics that define time/activity patterns (*26*); and environmental factors such as seasonality and community/area effect (*27*) all affect personal exposure to air toxics.

The main aim of the Measurement and Modeling of Air Toxic Concentrations for Health Effect Studies (MATCH) study was to help advance our understanding of the causes and magnitude of exposures to PAH and VOC and to establish whether collecting lifestyle information is sufficient to model PE reliably compared with exposures evaluated independently by personal samplers (*15*). To date, the modeling of PE to PAH has primarily focused on occupational exposures which tend to come from a single dominant PAH source (*28*). The challenge in modeling PE in nonoccupational settings lies in the fact that the exposure to PAH is complex due to their multitude of sources, both in indoor and outdoor microenvironments. This paper compares a number of models for 10 PAH using linear regression and supervised machine learning techniques and tests the models with an independent data set. Five models based on linear regression techniques were developed. Two further models predict PAH concentrations through the use of decision tree learning algorithms and neural networks respectively.

* Corresponding author phone: 00 44 121 414 3494; fax: 00 44 121 414 3709; e-mail: r.m.harrison@bham.ac.uk.
[†] University of Birmingham.
[‡] Faculty of Science, University of Malta.
[§] Faculty of ICT, University of Malta.

10.1021/es102529k

Published on Web 11/19/2010

2. Methodology

2.1. Recruitment of Subjects. For the details of the microenvironment and PE sampling campaigns the reader is referred to ref *29*. The subjects were recruited according to a matrix of determinants possibly effecting their PE, namely: 1) their geographical location subdivided as urban, suburban, and rural in London, Birmingham, and South Wales; 2) those who lived in close proximity to a trafficked road termed as a first line (FL) property and those who did not; 3) those who had a house with an integral garage (IG) or not; and 4) those who were exposed to Environmental Tobacco Smoke (ETS) or not. For further details on the recruitment strategy the reader is referred to ref *30*. The description of the data collected related to the subjects participating in MATCH apart from the air sampling is found in detail in ref *15*.

2.1. Sampling and Analytical Methods. In summary, 100 healthy adult volunteers participated in this project, and one 24 h integrated PE sample was collected per subject, using active sampling on a 47 mm quartz fiber filter. The samples were then stored at −20 °C, extracted by solvent extraction, and analyzed by gas chromatography−mass spectrometry (GC-MS). For further details of the analytical methods, the reader is referred to ref *31*.

2.2. Data Collection. A thirty-minute resolution time activity diary (TAD) was recorded by each subject accompanying the PE sample. Apart from recording the location where the subjects were, they were asked to note down what activities they were performing, if there was any ventilation when residing indoors, indicating if doors or windows were open, and if they were exposed to ETS. If the subjects traveled to other places apart from their home they were asked to indicate the places visited and the routes they had taken. Microenvironment concentrations (residential indoors, outdoor, transport, and workplaces) were also measured and are discussed in ref *32*. Wherever possible, samples for each microenvironment were collected during the morning rush hour and in the afternoon, once in summer and once in winter. During the microenvironment sampling the average ambient temperature and relative humidity were recorded. While sampling, further details on weather and traffic conditions, any ETS close to the sampler and other details that might have been useful in understanding the PAH sources were recorded in an Information Sheet. The Sampling Questionnaire gathered information on activities carried out involving the use of solvents or generation of particles in the house or the immediate vicinity that affected the PE on sampling day. The ETS Questionnaire contained detailed information on the number of smokers, number of cigarettes smoked, indoors or outdoors, and at which distance from the sampler these were smoked. Further details on the ventilation modes in operation in the enclosed space were also noted down. The Storage Questionnaire served to have information on possible PAH sources in the integral garage of a house. The Location Sheet Questionnaire identified the route the subjects took to go to places out of the house and their proximity to roads. The indoor characteristics of these places, such as their location compared to street level and any possible sources related to redecoration and smoking, were also recorded. These questionnaires are available in Appendix 3 of ref *30*.

2.3. Model Development. The models developed to predict the PE to PAH were based on measured data in various microenvironments and TADs. Two independent, measured PE data sets were chosen to develop and validate the PE model. 75% of the data set (68 samples) was used to develop the model and hereinafter will be termed the training data set, and the remainder 25% (23 samples) was kept to validate it, referred to as the testing data set. The remainder, 9% of the collected samples (9 samples), was used in the analytical method development and hence not available for modeling purposes. For the model development and validation, the data were chosen randomly (refer to Table S1) but ensuring an approximately equal proportion of ETS-exposed subjects in each data set based upon the corresponding 3-ethenylpyridine (3-EP) concentrations. 3-EP is an ETS marker that was collected concurrently with PAH. Data were analyzed using SPSS 15 for Windows (SPSS Inc.1989−2006, Release 15.0.0). As the PAH data distributions were generally positively skewed, log transformed data were used in all statistical analyses.

Using the collected data, various models have been developed to predict PE to PAH. A summary of the models developed is given hereunder.

i. Linear Regression Model - Model 1. Model 1 assessed associations between PE and microenvironment concentrations measured in homes and workplaces directly related to every subject. The equations used in Model 1 were

$$\log(Y_i) = \alpha + \beta^*\log(X_{i,\text{home}}) + \varepsilon_i \quad (1)$$

$$\log(Y_i) = \alpha + \beta^*\log(X_{i,\text{workplace}}) + \varepsilon_i \quad (2)$$

where Y_i is the measured personal exposure for a subject i, α and β are the intercept and the slope, respectively, $X_{i,\text{home}}$ is the average home concentration for the subject i, and $X_{i,\text{workplace}}$ is the workplace concentration for subject i. The terms ε_i represent the random error.

ii. Time Activity Weighted Model - Model 2. This model predicted the PE by summing up every time fraction spent in each microenvironment multiplied by the concentration of each microenvironment visited as shown in eq 3

$$P_{ij} = \sum \frac{t_{ijk}{}^* X_{ik}}{T_{ij}} \quad (3)$$

where P_{ij} is personal predicted exposure for a subject i on a day j, t_{ijk} is the time spent in microenvironment k by subject i on a day j, X_{ik} is the concentration, representative of microenvironment k for subject i, and T_{ij} is the total time spent in all different microenvironments for subject i on a day j, which is the same as the sampling time for subject i on a day j.

The microenvironment concentrations used in Model 2 for homes and workplaces were the data collected directly in each subjects' home and workplace. That is 21 workplaces and 59 home microenvironments for PAH. For those subjects where no data for home or work were available and for all the rest of the microenvironments (streets, transport and other indoor microenvironments) an average concentration representative of each microenvironment was used (Supporting Information - Table S2).

iii. Time Activity Weighted Model - Model 3. This predicted the PE by summing up the time fraction spent in each microenvironment multiplied by the concentration of each microenvironment visited as described by eq 3. In this model, the microenvironment concentrations used for homes and workplaces were not the actual collected data for each subject's home and workplace but an average value, representative of the microenvironment. Thus eq 3 was slightly modified to

$$P_{ij} = \sum \frac{t_{ijk}{}^* X_k}{T_{ij}} \quad (4)$$

where X_k is the concentration representative of microenvironment k. For this purpose a detailed list of stratified microenvironment concentrations has been developed for all the microenvironments, namely homes, workplaces, streets, transport, and other indoor microenvironments taking into account and applying different strata such as

location, season, time of the day traffic exposure, and ETS exposure (Supporting Information - Table S3).

iv. Hybrid Model - Model 4. Model 4 predicted the PE by summing up the time fraction spent in each microenvironment multiplied by the concentration of each microenvironment visited and accounting also for external factors that might affect exposure as add-on variables

$$Y_{ij} = \alpha \cdot P_{ij} + \sum \beta_m A_m + \sum \gamma_n F_n \quad (5)$$

where Y_{ij} is the observed personal exposure for a subject i on a day j, P_{ij} is the predicted personal exposure for a subject i on a day j as calculated in Model 3, α is the coefficient associated with personal exposure P_{ij}, A_m are different explanatory variables describing activities performed on a day j by a volunteer i or characteristics associated to a volunteer i, β_m is the coefficient associated with the explanatory variable A_m, F_n represents the time spent in doing different activities, and γ_n is the coefficient associated with the factor F_n.

The explanatory variables, A_m, associated with activities were extracted from the TAD, the ETS, and Activity Questionnaires. The explanatory variables related to characteristics were extracted from the Sampling, ETS, and Storage and Location Sheet questionnaires. The explanatory variables have a value of 1 if the activity is performed or the characteristic is present, and 0 if not. The variables representing the time associated with different variables, F_n, were measured in minutes and were extracted from the TAD and the Location Sheets for Traveling. A total of 63 add-on variables (an example of these add-ons would be Time in ETS, entering the value in minutes would indicate the time that the subject was exposed to smoke), A_m and F_n, extracted from the available information were included into the model (Supporting Information - Table S4).

The PE models were developed in three steps. Initially, using a stepwise option, the linear regression model identified the optimum number of variables to get the highest correlation coefficient. The criterion followed to enter new variables was to have a probability of F less than or equal to 0.05. The criterion to remove variables was if the probability of F was greater than or equal to 0.10. In the second stage the variables selected automatically by the statistical package were checked to have a scientific meaning and that the least number of selected variables explained most of the variation. The final linear regression model would enter the selected variables from the previous stage.

The best model developed with the training data set was used to predict concentrations in the testing data set, and the most important variables for each compound are summarized in Table S5.

v. Univariate General Linear Model (UGLM) - Model 5. To check if the Model 4 could be improved, a Univariate General Linear Model (UGLM) was used. The predicted concentration was related to a number of categorical (fixed) variables (for example Summer, so if a value for this variable is 1 it indicates the sample was taken in summer and the contrary if 0) and other covariates (these are continuous predictor variables for example Time-in-a-car which would indicate the accumulation of time a subject spends in a car). All the original variables drawn up for the PAH model were entered, and a univariate analysis of variance was carried out, each time removing the variable with the least significance. This removal procedure was done repeatedly until the variables remaining were statistically significant ($p < 0.05$), and thus the GLM parameters were obtained. With reference to Table S6, the model parameters for every PAH were extracted and computed the predicted concentrations with the training data set. The R^2 of this approach shows that this parsimonius model was not an improvement compared to the Model 4 when run for the testing data set. However it is notable that Model 5 enters variables into the model which are different from the Model 4.

vi. Machine Learning Techniques - Decision Trees - Model 6. Apart from the models generated through linear regression, the applicability of machine learning techniques to model PE to PAH concentrations were investigated. Tree learning methods allow for a clear and comprehensive class description to be learned from the provided samples. Through the calculation of the information gain, the most important parameters that affect the final PAH concentrations can be determined from the top root nodes of the tree.

Various decision tree learning algorithms have been proposed. Following work done by Hunt in the late 1950s and early 1960s, Quinlan continued to improve on the developed techniques and released the Iterative Dichottomizer 3 (ID3) and the improved C4.5 decision tree learners (*33*). In this study the C4.5 decision tree algorithm was used.

Ten PAH are treated independently, and separate data sets each consisting of the 40 time-values corresponding to the duration spent by each participant in the various microenvironment as well as the 63 other add-on boolean parameters were initially generated and are tabulated in Table S7. Since the C4.5 algorithm is a classification technique, the measured concentrations were also binned into discretized sets covering 6 categories in the following concentration ranges: below method detection limit -0.10 ng/m^3 (Category A), 0.10$-$0.12 ng/m^3 (Category B), 0.12$-$0.20 ng/m^3 (Category C), 0.20$-$0.25 ng/m^3 (Category D), 0.25$-$1.00 ng/m^3 (Category E), and more than 1.00 ng/m^3 (Category F).

The European fourth air quality daughter Directive 2004/107/EC, relating to PAH in ambient air, specifies a target value for B(a)P of 1 ng m^{-3} (annual average) to be achieved by 2012. In the UK, the Expert Panel on Air Quality Standards (EPAQS) has recommended an Air Quality Standard of 0.25 ng m^{-3} B(a)P as an annual average. Based on these levels, the classes A to F proposed above were determined, and each sample was allocated to a nominal class.

Using the independent test data set, classification accuracy describes the percentage of correctly classified testing samples in the respective bins. To test classification accuracy, 10-fold cross-validation was used. This involves splitting the data into ten sets from which nine sets are used to train the classifier and the remaining set is used to test the inferred tree. Repeating this process for nine other times will allow each sample to be treated as a test case at least once. The average accuracy was then computed after determining the number of samples for which the classifier gave the expected class.

vii. Neural Networks - Model 7. Another machine learning technique that mimics the biological learning processes occurring in the human brain was used. Neural networks present a robust way to predict real-value concentrations after learning from a supplied sample set. Such networks connect a number of individual elements each of which take a set of inputs and produce a single real number. The learning algorithm determines numeric weights to apply between each of these neurons to obtain the desired output. One main advantage of this technique is that it can produce good results even when supplying it with noisy and incomplete data.

A two-layer feed-forward network with sigmoid hidden neurons and linear output neurons was fitted to the generated data sets. The network was trained using the Levenberg$-$Marquardt back-propagation algorithm. Twenty neurons were set in the hidden layer. The generated data sets for each compound representing 40 time values corresponding to the duration spent by each participant in the various microenvironment as well as the 63 other add-on boolean parameters (Table S7) were again used. As with the linear models, 75% of the data were used to train the classifier, while 25% of the

TABLE 1. Model Performance As Expressed by the Coefficient of Determination (R^2) for Models 1−5 and 7[a]

	model 1		model 2	model 3	model 4		model 5		model 7	
compound	home ($N = 36$) R^{2b}	work ($N = 20$) R^{2b}	($N = 68$) R^{2b}	($N = 68$) R^{2b}	training ($N = 68$) R^{2b}	testing ($N = 23$) R^{2b}	training ($N = 68$) R^{2b}	testing ($N = 23$) R^{2b}	training ($N = 68$) R^{2b}	testing ($N = 23$) R^{2b}
pyrene	0.226	0.002	0.066	**0.147**	**0.247**	**0.857**	**0.252**	0.008	**0.450**	**0.763**
benzo(a)anthracene	0.184	0.009	0.024	0.015	**0.661**	**0.261**	**0.695**	0.008	**0.752**	**0.714**
chrysene	**0.361**	0.191	0.006	0.049	**0.334**	0.073	**0.417**	0.001	**0.640**	**0.700**
benzo(b)fluoranthene	0.121	**0.502**	0.000	0.023	**0.278**	**0.360**	**0.274**	0.001	**0.509**	**0.706**
benzo(k)fluoranthene	0.023	0.131	0.003	0.022	**0.303**	0.185	**0.252**	0.032	**0.502**	**0.813**
benzo(a)pyrene	0.107	0.194	0.000	0.040	**0.346**	0.185	**0.393**	0.019	**0.567**	**0.640**
indeno(1,2,3-cd)pyrene	0.068	0.307	0.000	0.032	**0.282**	0.148	**0.165**	0.003	**0.434**	**0.797**
dibenzo(a,h)anthracene	**0.434**	0.009	0.003	0.042	**0.575**	**0.262**	**0.569**	0.001	**0.602**	**0.820**
benzo(ghi)perylene	0.076	**0.378**	0.000	0.084	**0.259**	0.122	**0.123**	0.044	**0.484**	**0.598**
coronene	0.008	0.250	0.001	0.101	**0.367**	**0.387**	**0.435**	0.012	**0.486**	**0.717**

[a] Bold values represent significant correlation values at the $p \leq 0.01$ level. [b] Statistic.

samples were left for testing. The correlation coefficient between the measured data and the predicted modeled data as predicted by a trained neural network was calculated for each component. This made possible direct comparisons with the linear regression model accuracies.

3. Results and Discussion

Descriptive statistics of the whole PAH measured data set, the training and the testing data sets appear in Table S1 and confirm that there was no bias in their choice. The results of model development appear in Table 1.

Viewed across the full suite of compounds, neither Model 1 nor Model 2 performed well. The time-weighted Model 3 which used average stratified data for all the cases explained less variability in personal exposure than the previous models. This was a consequence of the difficulties encountered in adequately stratifying home microenvironments with the relatively small number of samples collected. Even when the total sample size is fairly big ($N = 68$ samples), the large number of different strata to account for integral garage, ETS exposure, first line properties or location within a city, reduces considerably the sample size per stratum, even down to just 2 samples per stratum. The range of activities that the subjects were engaged to in their normal life was reflected in the specific home and workplace levels. The stratified PAH levels, representative of each different stratum of exposure, did not reflect the various activities. To solve this difficulty, the hybrid Model 4 used the concentrations calculated in Model 3 and included a number of add-on variables that represented activities or home characteristics, that could not be reflected in the stratified data. None of the two proposed time-weighted models (Model 2 and Model 3), which only consider time spent in different microenvironments, could adequately explain the variance of the PAH compounds. However, Model 4 (Table 1), with the inclusion of add-ons was able to explain 25−65% of the variance of the PAH compounds (35% of variance for B(a)P). This model also performed relatively well in predicting VOC exposures (*15*). In the testing data set of Model 4, most R^2 values were lower (Table 4). The variables entered in Model 4 are described in the Supporting Information, Table S4 and in Table S5.

When the performance of Model 4 was tested as to whether predicted PE was above or below a threshold value in the validation data set, the model was successful in correctly classifying cases in the lower exposure band showing 83% to 100% of correctly classified cases (Table S8). However, the success rate decreased substantially when classifying cases in the upper exposure band, for many compounds failing totally to identify cases above the threshold. This is, however, not unexpected as all the models showed a tendency to underestimate the PE levels.

Results from Model 4, illustrated in Figure S2(a), show that intercepts were generally close to zero in the training data set, but the testing data set showed larger intercepts (Figure S2(b)). Similarly, from the scatter plots of Figure S2(a) and (b), correlating measured with predicted concentrations, the model generally slightly underestimated the values in the training data set and even more in the testing data set.

The model parameters entered and the corresponding R^2 values for Model 5 are summarized in Table S6. Similar or slightly higher values of R^2 compared to Model 4 were obtained for Pyr, B(a)A, Chry, B(a)P, D(a,h)A, and Cor (Table 1). Two variables, the number of cigarettes within 2 m from the sampler (covariate) and the identification if the sample was collected in summer (fixed variable), are significant for almost all the PAH in both model approaches. Nevertheless, Model 5 when run with the testing data set performed worse than Model 4.

In the case of Model 6, Figure 1 shows an example of an inferred decision tree for B(a)P, and the results for all compounds are presented in Table 2. The accuracy values were calculated by calculating how many modeled concentrations were classified in the correct bin. One main disadvantage of this approach is that if the correct value is very close to the upper or lower threshold of the bin and the predicted value surpasses the threshold by a very small amount, the sample is still considered to be incorrectly classified. Another disadvantage is that the accuracy values cannot be directly compared to the correlation coefficient computed for the linear models (Models 1−5). While the R^2 value indicates how well the predicted values match the measured concentrations, the accuracies presented here indicate how precise the algorithm is in determining the correct class. Figure S3 shows the number of concentrations modeled and measured in each category for B(a)A, B(a)P, and D(a,h)A, respectively.

It is interesting to note that for the majority of the compounds, the parameter representing the number of cigarettes within 2 m was designated as the root of the tree indicating a high information gain and hence a high characterizing factor to the modeled PE. Parameters designated in first level nodes included those representing time spent in a car, time spent in pub/bar/social club, and whether the car driven had a diesel engine. Such attributes were also found to have a high characterizing factor in the linear models and hence have a significant effect on the PAH.

In Model 7, the resulting correlation coefficients obtained after plotting the measured (expected) values versus the predicted (modeled) results as outputted by the neural network for the training set and the testing set are shown in Table 1. Figure S4 shows the plots obtained for Pyr, B(a)P, and D(a,h)A for modeled and measured concentrations. The

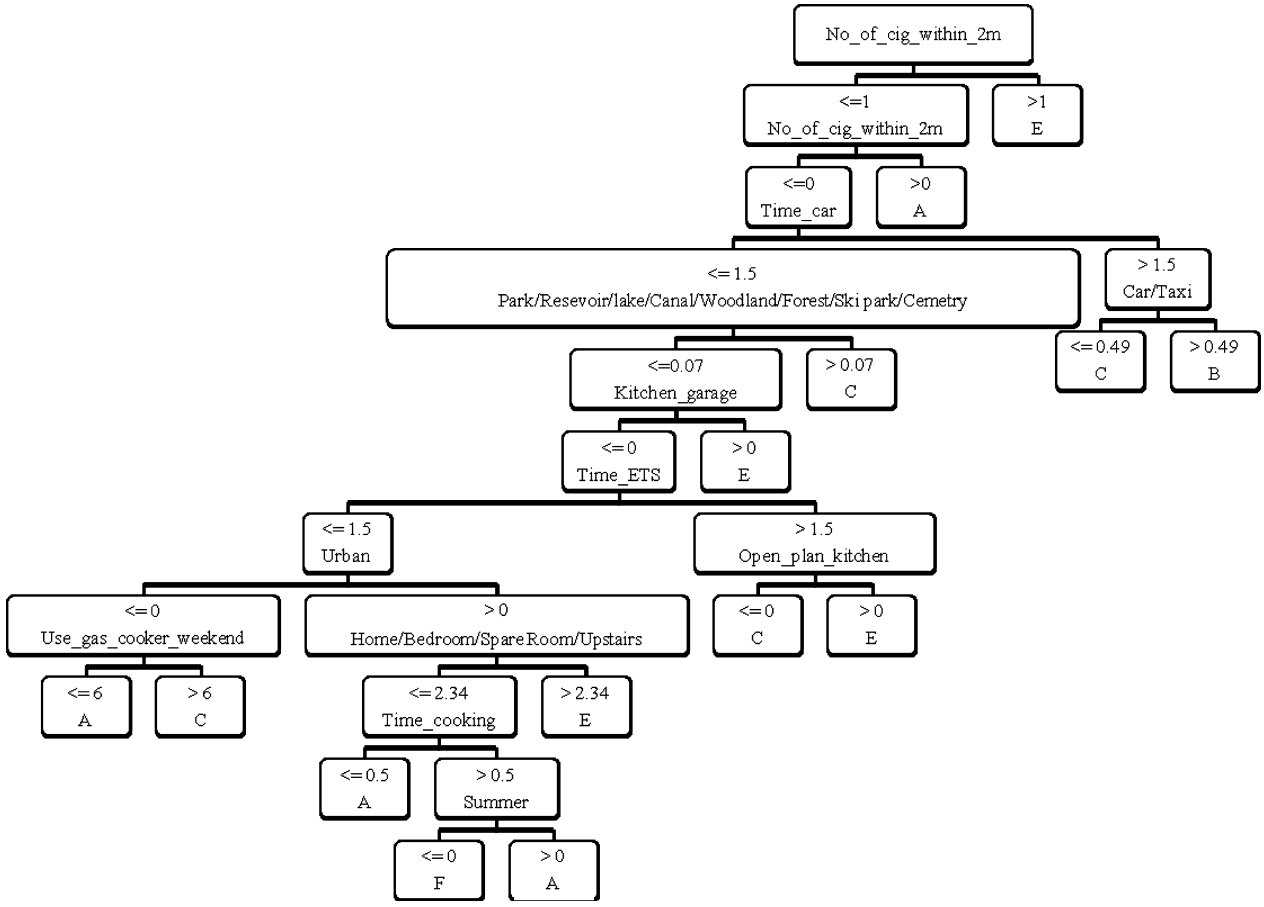

FIGURE 1. Decision tree for B(a)P determined by Model 6.

TABLE 2. Classification Accuracies of the Decision Tree Learning Algorithm (Model 6)

compound	accuracy/%
benzo(a)anthracene	69.4
benzo(a)pyrene	52.3
benzo(b)fluoranthene	36.4
benzo(ghi)perylene	27.5
benzo(k)fluoranthene	36.7
chrysene	30.7
coronene	39.8
dibenzo(a,h)anthracene	89.0
indeno(1,2,3-cd)pyrene	33.0
pyrene	18.3

strong correlation between the measured and predicted values achieved by Model 7 can be clearly seen in Figure S4. Using this model, an increase in the calculated R^2 for all compounds is noted in Table 1.

The evolution of the linear regression Models 1−5 indicated that for this technique the hybrid Model 4 and the UGLM Model 5 were the best approaches. An important difference between Model 4 and Model 5 is that the models identified different variables as determining the predicted PAH concentration. Models 4 and 5 identify as add-on variables those related to ETS (e.g., number of cigarettes within 2 m), other combustion (e.g., incense burning), and traffic (e.g., use of bus). The same sources have been widely associated with PAH in the literature (*1, 34, 35*).

The two machine learning techniques tested performed relatively well. As indicated in Table 2, Model 6 produced some high classification accuracies. However, these cannot be directly compared with the correlation coefficient results produced by the linear regression methods and neural networks. The cigarettes parameter used in several of the predictive equations of Model 5 (Table S6) is identified by Model 6 as the most significant, implying it is a crucial factor influencing PE in nonoccupational settings.

The neural network results from Model 7 show significant improvements on the outputs of Models 4 and 5 as shown in Table 1 when dealing with the training data set and when modeling PE from unseen samples (testing data set). This good performance motivates further investigation to exploit the potential of machine learning techniques for such modeling and the use of support vector machines, Bayesian statistics, and component analysis. Although the main limitation was the relatively small number of samples to build the microenvironment database and identify the factors which are determinants of PE, the authors have confidence that with larger data sets including more highly exposed subjects, the hybrid model and the machine learning algorithms will be able to predict PE more accurately. However, and despite the small sample size, the main sources affecting personal exposure to PAH, such as exposure to ETS and traffic, have been successfully identified by several different model approaches, enhancing our confidence in the proposed models. These main PAH sources are incorporated into the models using a reduced set of add-on variables, identified for each compound, which can be easily collected from subjects using questionnaires and time activity diaries. Therefore, using lifestyle information collected in questionnaires, these models are capable of predicting PAH exposures, which can be subsequently used in epidemiological studies and to assess health risk associated with exposures.

Deterministic models enable pollutant concentrations to be modeled in a more dynamic way by taking into account

the space-time interactions that characterize pollution processes in air or otherwise in order to use them to predict PE. These models can be complementary to our empirical models which use airborne PAH concentrations as input to predict PE levels. Alternatively, while empirical models often do not reflect the physical/chemical phenomena that cause the exposures, they implicitly reflect the interdependencies within the measured dependent and independent variables, regardless of whether they are known or considered by the modeler, making them suitable to be applied to other locations and populations.

Finally, the terms in our models are specific to the data set from which they have been calculated and hence caution should be applied when extrapolating the proposed models (i.e., selected variables and coefficients) to different geographical regions, countries, times, climates, and locations with markedly different sources of pollutants (*15*). However, the different modeling approaches proposed in this study are suitable for application elsewhere, with the hybrid, and machine learning technique models being recommended for PAH exposures.

Acknowledgments

The authors acknowledge financial assistance from the Health Effects Institute in the data collection and development of Models 1−4. This research was in part conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the United States Environmental Protection Agency (U.S. EPA) (Assistance Award No. 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 U.S. EPA or motor vehicle and engine manufacturers.

Supporting Information Available

Tables S1-S8 and Figures S1-S4. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Harrison, R. M.; Smith, D. J. T.; Luhana, L. Source apportionment of atmospheric polycyclic aromatic hydrocarbons collected from an urban location in Birmingham, UK. *Environ.l Sci. Technol.* **1996**, *30*, 825–832.

(2) Brandt, H. C. A.; Watson, W. P. Monitoring human occupational and environmental exposures to polycyclic aromatic compounds. *Ann. Occup. Hyg.* **2003**, *47*, 349–378.

(3) Benner, B. A.; Gordon, G. E.; Wise, S. A. Mobile sources of atmospheric polycyclic aromatic hydrocarbons - a roadway tunnel study. *Environ. Sci. Technol.* **1989**, *23* (10), 1269–1278.

(4) Binet, S.; Pfohl-Leszkowicz, A.; Brandt, H.; Lafontaine, M.; Castegnaro, M. Bitumen fumes: Review of work on the potential risk to workers and the present knowledge on its origin. *Sci. Total Environ.* **2002**, *300*, 37–49.

(5) Chiang, T.-A.; Wu, P.-F.; Ko, Y.-C. Identification of carcinogens in cooking oil fumes. *Environ. Res.* **1999**, *81*, 18–22.

(6) Smith, K. R. *Biofuels, Air Pollution, and Health - A Global Review*; Plenum Press: New York, 1987.

(7) Sakai, R.; Siegmann, H. C.; Sato, H.; Voorhees, A. S. Particulate matter and particle-attached polycyclic aromatic hydrocarbons in the indoor and outdoor air of Tokyo measured with personal monitors. *Environ. Res.* **2002**, *89* (1), 66–71.

(8) Schauer, J. J.; Kleeman, M. J.; Cass, G. R.; Simoneit, B. R. T. Measurement of emissions from air pollution sources. 3. C-1-C-29 organic compounds from fireplace combustion of wood. *Environ. Sci. Technol.* **2001**, *35*, 1716–1728.

(9) Oanh, N. T. K.; Reutergardh, L. B.; Dung, N. T. Emission of polycyclic aromatic hydrocarbons and particulate matter from domestic combustion of selected fuels. *Environ. Sci. Technol.* **1999**, *33*, 2703–2709.

(10) Rogge, W. F.; Hildemann, L. M.; Mazurek, M. A.; Cass, G. R.; Simoneit, B. R. T. Sources of fine organic aerosols. 5. Natural Gas Home Appliances. *Environ. Sci. Technol.* **1993**, *27*, 2736–2744.

(11) Heudorf, U.; Angerer, J. Internal exposure to PAHs of children and adults living in homes with parquet flooring containing high levels of PAHs in the parquet glue. *Int. Arch. Occup. Environ. Health* **2001**, *74*, 91–101.

(12) Chuang, J. C.; Mack, G. A.; Kuhlman, M. R.; Wilson, N. K. Polycyclic aromatic hydrocarbons and their derivatives in indoor and outdoor air in an 8-home study. *Atmos. Environ.* **1991**, *25*, 369–380.

(13) Naumova, Y. Y.; Eisenreich, S. J.; Turpin, B. J.; Weisel, C. P.; Morandi, M. T.; Colome, S. D.; Totten, L. A.; Stock, T. H.; Winer, A. M.; Alimokhtari, S.; Kwon, J.; Shendell, D.; Jones, J.; Maberti, S.; Wall, S. J. Polycyclic aromatic hydrocarbons in the indoor and outdoor air of three cities in the US. *Environ. Sci. Technol.* **2002**, *36*, 2552–2559.

(14) Belpomme, D.; Irigaray, P.; Hardell, L.; Clapp, R.; Montagnier, L.; Epstein, S.; Sasco, A. J. The multitude and diversity of environmental carcinogens - A Review. *Environ. Res.* **2007**, *105*, 414–429.

(15) 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.

(16) Leung, P. L.; Harrison, R. M. Evaluation of personal exposure to monoaromatic hydrocarbons. *Occup. Environ. Med.* **1998**, *55*, 249–257.

(17) Payne-Sturges, D. C.; Burke, T. A.; Breysse, P.; Diener-West, M.; Buckley, T. J. Personal exposure meets risk assessment: A comparison of measured and modeled exposures and risks in an urban community. *Environ. Health Perspect.* **2004**, *112*, 589–598.

(18) Liu, W.; Zhang, J.; Korn, L. R.; Zhang, L.; Weisel, C. P.; Turpin, B.; Morandi, M.; Stock, T.; Colome, S. Predicting personal exposure to airborne carbonyls using residential measurements and time/activity data. *Atmos. Environ.* **2007**, *41*, 5280–5288.

(19) Serrano-Trespalacios, P. I.; Ryan, L.; Spengler, J. D. Ambient, indoor and personal exposure relationships of volatile organic compounds in Mexico City metropolitan area. *J. Expo. Anal. Environ. Epidemiol.* **2004**, *14*, S118–S132.

(20) Wilson, J. G.; Kingham, S.; Pearce, J.; Sturman, A. P. A review of intraurban variations in particulate air pollution: Implications for epidemiological research. *Atmos. Environ.* **2005**, *39*, 6444–6462.

(21) Dodson, R. E.; Houseman, E. A.; Levy, J. I.; Spengler, J. D.; Shine, J. P.; Bennett, D. H. Measured and modeled personal exposures to and risks from volatile organic compounds. *Environ. Sci. Technol.* **2007**, *41*, 8498–8505.

(22) Edwards, R. D.; Jurvelin, J.; Saarela, K.; Jantunen, M. VOC concentrations measured in personal samples and residential indoor, outdoor and workplace microenvironments in EXPOLIS—Helsinki, Finland. *Atmos. Environ.* **2001**, *35*, 4531–4543.

(23) Kim, Y. M.; Harrad, S.; Harrison, R. M. Levels and sources of personal inhalation exposure to volatile organic compounds. *Environ. Sci. Technol.* **2002**, *36*, 5405–5410.

(24) Loh, M. M.; Houseman, E. A.; Gray, G. M.; Levy, J.I.; Spengler, J. D.; Bennett, D. H. Measured concentrations of VOCs in several non-residential microenvironments in the United States. *Environ. Sci. Technol.* **2006**, *40*, 6903–6911.

(25) Harrison, R. M.; Thornton, C. A.; Lawrence, R. G.; Mark, D.; Kinnersley, R. P.; Ayres, J. G. Personal exposure monitoring of particulate matter, nitrogen dioxide, and carbon monoxide, including susceptible groups. *Occup. Environ. Med.* **2002**, *59*, 671–679.

(26) Edwards, R. D.; Schweizer, C.; Llacqu, V.; Lai, H. K.; Jantunen, M.; Bayer-Oglesby, L.; Künzli, N. Time-activity relationships to VOC personal exposure factors. *Atmos. Environ.* **2006**, *40*, 5685–5700.

(27) Sexton, K.; Adgate, J. L.; Ramachandran, G.; Pratt, G. C.; Mongin, S.;J.; Stock, T. H.; Morandi, M. Comparison of personal, indoor, and outdoor exposures to hazardous air pollutants in three urban communities. *Environ. Sci. Technol.* **2004**, *38*, 423–430.

(28) Burstyn, I.; Kromhout, H.; Kauppinen, T.; Heikkila, P.; Boffetta, P. Statistical modelling of the determinants of historical exposure to bitumen and polycyclic aromatic hydrocarbons among paving workers. *Ann. Occup. Hyg.* **2000**, *44*, 43–56.

(29) 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.

(30) 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: Boston, MA, 2009.
(31) Delgado-Saborit, J. M.; Aquilina, N.; Baker, S.; Harrad, S.; Meddings, C.; Harrison, R. M. Determination of atmospheric particulate-phase polycyclic aromatic hydrocarbons from low volume air samples. *Anal. Methods* **2010**, *2*, 231–242.
(32) Aquilina, N. *Evaluation of human exposure to airborne carcinogenic compounds*, Ph.D. Thesis, University of Birmingham, UK, 2009.
(33) Kohavi, R.; Quinlan, R. Decision Tree Discovery, 1990.
(34) Lim, L. H.; Harrison, R. M.; Harrad, S. The contribution of traffic to atmospheric concentrations of polycyclic aromatic hydrocarbons. *Environ. Sci. Technol.* **1999**, *33*, 3538–3542.
(35) Levy, J. I.; Houseman, E. A.; Spengler, J. D.; Loh, P.; Ryan, L. Fine particulate matter and polycyclic aromatic hydrocarbon concentration patterns in Roxbury, Massachusetts: A community-based GIS analysis. *Environ. Health Perspect.* **2001**, *109*, 341–347.

ES102529K