

Characterizing the Relationship between Personal Exposures to VOCs and Behavioral, Socioeconomic, Demographic Variables: Analysis of NHANES VOC Project Data Set

Y. Yan, S.W. Wang, and P.G. Georgopoulos • Computational Chemodynamics Laboratory (www.ccl.rutgers.edu)

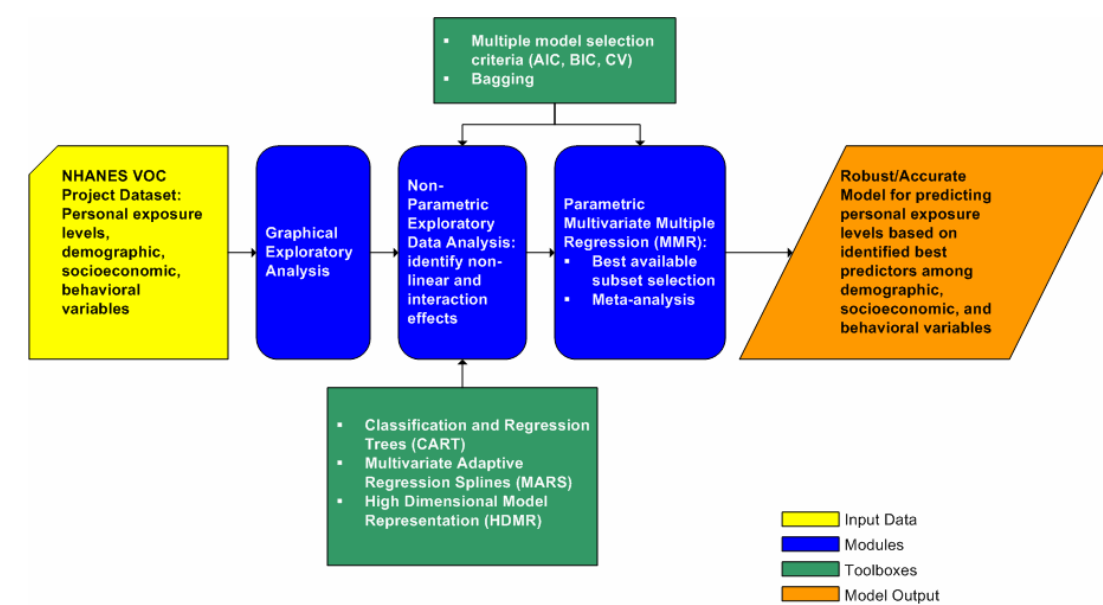
Environmental and Occupational Health Sciences Institute, a Joint Institute of UMDNJ-Robert Wood Johnson Medical School and Rutgers University, Piscataway, NJ



ABSTRACT

This study presents the application of a systematic data analysis framework consisting of graphical exploratory analysis, canonical correlation analysis, Classification And Regression Trees (CART), and Multivariate Adaptive Regression Splines (MARS) to identify “best predictors” of personal exposures to VOCs using data collected in the 1999-2000 NHANES VOC Project. These statistical techniques are employed to address limitations and challenges in the complex NHANES VOC dataset, such as missing values, collinearity, nonlinearity, interaction effects etc. This dataset contains the measurements of personal exposures to 10 VOCs for 659 subjects between the ages of 20 and 59 years. Data on individual demographic and socioeconomic status, as well as time and activity patterns for the exposure period are also available for these subjects. The data analysis outcomes provide valuable information for identifying significant exposure factors among demographic, socioeconomic, and activity variables that affect personal exposures to VOCs.

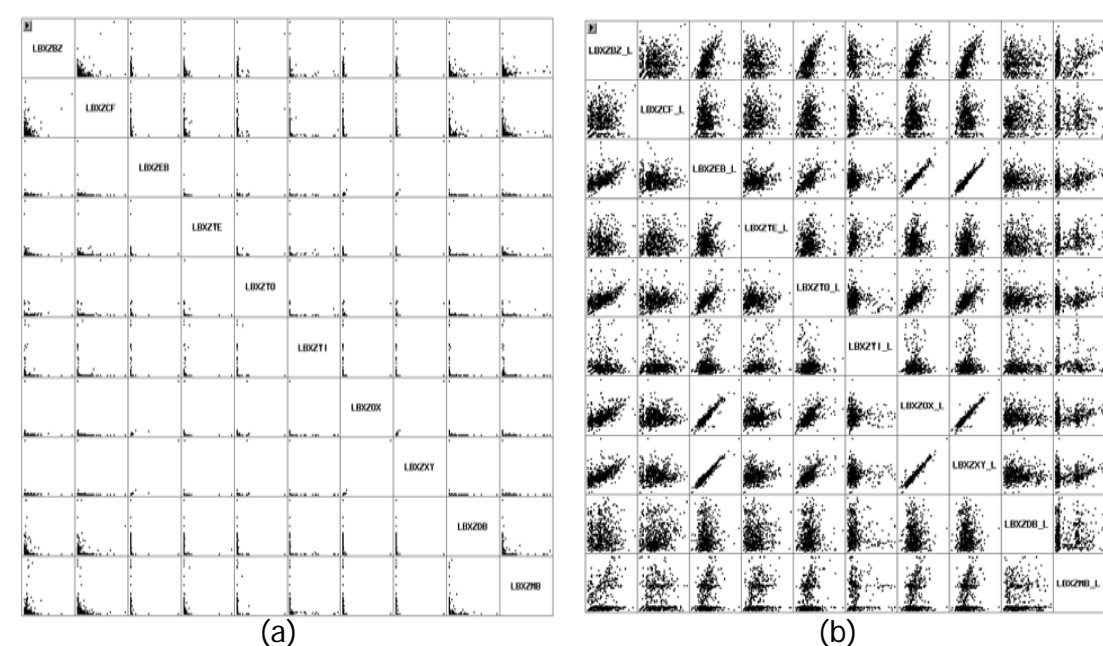
METHODOLOGY



A systematic data analysis framework for identifying “best predictors” among environmental, demographic, and activity variables for determining personal exposure levels, utilizing a suite of statistical and data mining techniques.

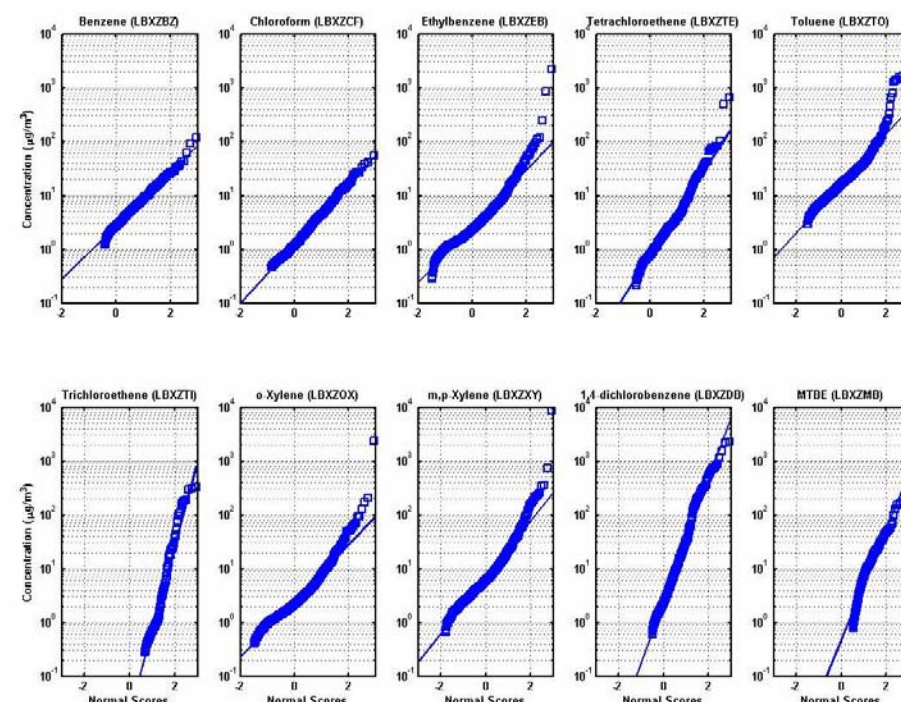
RESULTS

Graphical Exploratory Analysis



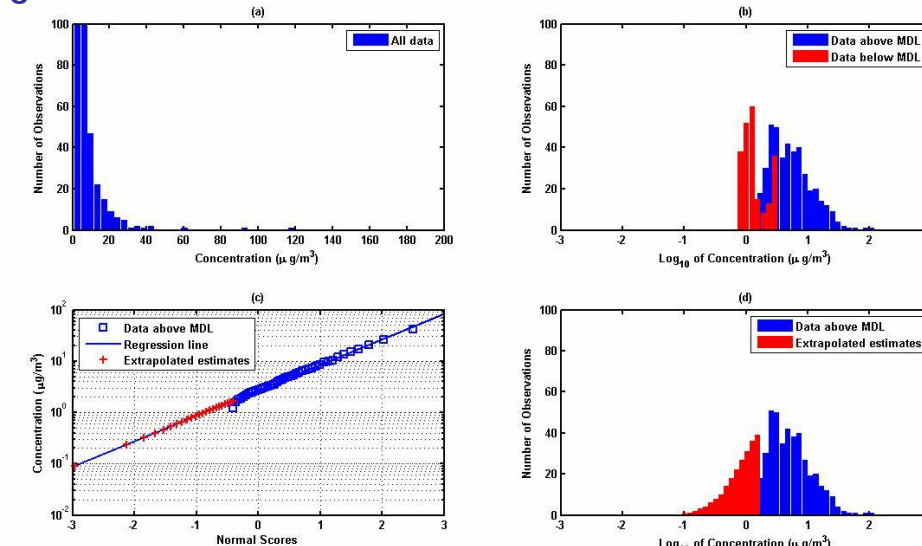
The pair-wise scatter plots of personal exposure levels for the 10 VOCs using (a) raw data (b) log-transformed. Figure (b) shows that there are significant linear associations among LBXZEB (Ethylbenzene), LBXZOX (o-Xylene), and LBXZXY (m,p-Xylene).

Identifying outliers



For identifying the possible outliers, normal probability plots were generated for the 10 personal VOCs concentrations (using data above detection limits). Several data points appear to be away from the majority of the distribution for the cases of ethylbenzene, toluene, o-xylene, m,p-xylene, and tetrachloroethene. Further, it was found that the first three highest values of ethylbenzene, o-xylene, and m,p-xylene personal concentrations are related to the same three subjects.

Handling Non-detects



Step-by-step application of a robust parametric method for handling non-detects in the Benzene personal air concentration. Panel (a) shows the histogram of the original reported Benzene personal air concentrations. Panel (b) shows the histogram of the Log transformed Benzene personal air concentrations. Panel (c) shows the regression results of using the robust parametric method for fitting the data above the detection limit in normal probability plot. Panel (d) shows the histogram generated by combining data above the detection limit with the extrapolated estimates of the non-detects.

Canonical Correlation Analysis

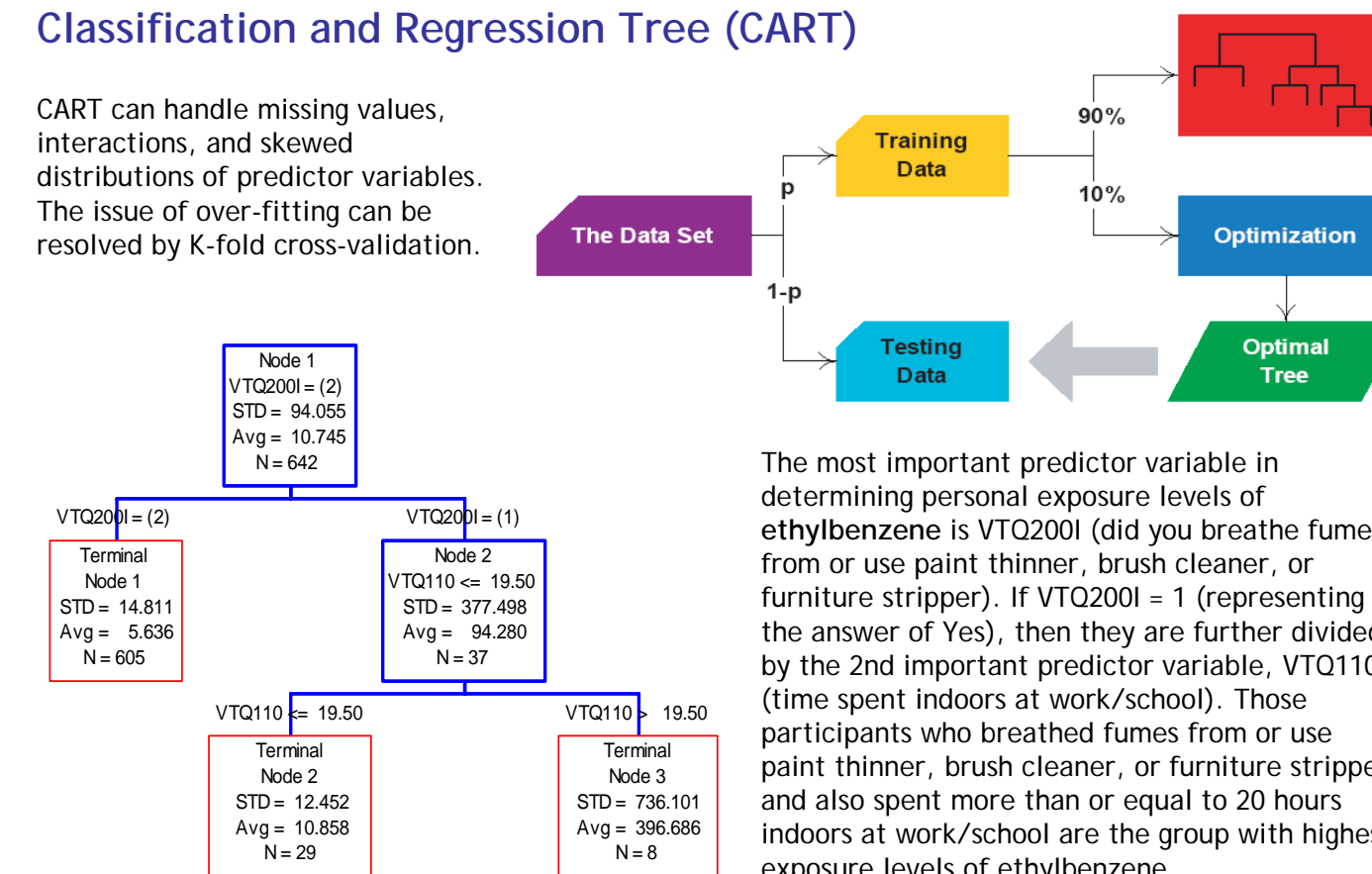
Canonical correlation analysis was conducted to reveal the extent of correlations among predictor and response variables. This was achieved by finding two sets of variables comprised by the linear combinations of original variables, called canonical variates, having maximal correlation. It was found that o-xylene, m,p-xylenes, and ethylbenzene contribute significantly to the canonical variate of personal exposure levels. The other canonical variate (corresponding to the exposure factors) has significant contributions from 5 out of a total of 36 exposure factors.

No.	Code	Exposure Levels Variables (VOCs)	Coefficients
1	LBXZXY	m,p-Xylenes	1.1302
2	LBXZEB	Ethylbenzene	-0.7587
3	LBXZOX	o-Xylene	-0.7492

No.	Code	Exposure Factors Descriptions	Coefficients
1	VTQ200I	Did you breathe fumes from or use paint thinner, brush cleaner, or furniture stripper?	-0.4216
2	VTQ130	Pump gas into a car or motor vehicle?	-0.3692
3	DMD140	Education	0.3194
4	VTQ150	In drycleaning shop, drycleaned clothes?	0.3136
5	RIDRETH1	Race/Ethnicity	0.1823

Classification and Regression Tree (CART)

CART can handle missing values, interactions, and skewed distributions of predictor variables. The issue of over-fitting can be resolved by K-fold cross-validation.



The most important predictor variable in determining personal exposure levels of ethylbenzene is VTQ200I (did you breathe fumes from or use paint thinner, brush cleaner, or furniture stripper). If VTQ200I = 1 (representing the answer of Yes), then they are further divided by the 2nd important predictor variable, VTQ110 (time spent indoors at work/school). Those participants who breathed fumes from or use paint thinner, brush cleaner, or furniture stripper, and also spent more than or equal to 20 hours indoors at work/school are the group with highest exposure levels of ethylbenzene.

Optimal CART tree model constructed for the personal air concentration of ethylbenzene (LBXZEB)

Multivariate Adaptive Regression Splines (MARS) Analysis

MARS method automates all aspects of model development and model deployment for identifying an optimal model. The ‘optimal’ MARS model is the one with the lowest GCV (generalized cross-validation) measure. For determining the personal exposure levels of benzene, there are 7 basis functions (BF) constructed for the optimal MARS model.

Response Variables	Basis Functions	Models	R ² / GCV R ²
Benzene (LBXZBZ)	BF1 = (VTQ100 > .); BF3 = (VTQ100 = 2) * BF1; BF5 = (RIDRETH2 = 3 OR 4); BF7 = (VTQ040 > .) * BF1; BF9 = (VTQ040 = 1) * BF7; BF11 = (VTQ050 > .); BF13 = (VTQ050 = 1 OR 2) * BF11;	Y = 2.376 + 2.746 * BF3 + 2.724 * BF5 + 3.151 * BF9 + 2.129 * BF13;	0.078 / 0.046

BF1 = (VTQ100 > .), indicating the status of missing value for the variable VTQ100. If not missing, BF1 is equal to 1. If missing, BF1 is equal to 0. The basis functions of BF3, BF5, BF9, and BF13 were incorporated into the constructed MARS model. These four basis functions actually represents the identification of the corresponding four important predictor variables (VTQ100, RIDRETH2, VTQ040, and VTQ050) in determining personal exposure levels of benzene, where these variables represent the responses of the following questions:

- VTQ100: were any windows open in your home
- RIDRETH2: race/ethnicity
- VTQ040: home built less than 5 years ago
- VTQ050: description of street where you live

CONCLUSIONS

- Significant linear correlations were found among the following 5 personal air concentrations: benzene, ethylbenzene, toluene, o-xylene, and m,p-xylenes.
- Three possible outliers in the data points of personal VOCs concentrations were identified through the normal probability plots.
- Cross-comparison among Canonical correlation analysis, CART analysis, and MARS analysis reveals that the time-activity variables of VTQ200I (did you breathe fumes from or use paint thinner, brush cleaner, or furniture stripper), VTQ090 (hours spent indoors at home), and VTQ110 (hours spent indoors at work/school) coupled with demographic variables of INDHINC (household income) and RIDRETH2 (race/ethnicity) are the important predictors for the personal air concentrations of the BTEX family (benzene, toluene, ethylbenzene, o-xylene, and m,p-xylene).

Acknowledgements

This work has been funded in part by the Mickey Leland National Urban Air Toxics Research Center and by the USEPA-funded Center for Exposure and Risk Modeling (CERM), under Cooperative Agreement # CR-83162501. This work has not been reviewed by and does not represent the opinions of the funding agencies.