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, 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), LBX2OX (o-Xylene), and LBX2ZV (m,p-Xylene).

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 variables, having maximal correlation. It was found that o-xylene, m,p-xylene, 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.

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.

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

Multi-variate 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.

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 plots. Panel (d) shows the histogram generated by combining data above the detection limit with the extrapolated estimates of the non-detects.

Carter's tree model for predicting the personal exposure levels of benzene (LBXZEB)

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), LBX2OX (o-Xylene), and LBX2ZV (m,p-Xylene).

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 variables, having maximal correlation. It was found that o-xylene, m,p-xylene, 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.

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.

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

Multi-variate 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.

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 plots. Panel (d) shows the histogram generated by combining data above the detection limit with the extrapolated estimates of the non-detects.

CONCLUSIONS
Significant linear correlations were found among the following 5 personal air concentrations: benzene, ethylbenzene, toluene, o-xylene, and m,p-xylene.

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 VTQ001 (did you breathe fumes from or use paint thinner, brush cleaner, or furniture stripper), VTQ040 (hours spent indoors at work/school), and VTQ050 (describe of street where you live)

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.