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

Presented at the SRA 2006 Annual Conference Baltimore, MD • December 4-6, 2006

by
S.W. Wang, S. Tong, P. G. Georgopoulos

Computational Chemodynamics Laboratory
Environmental and Occupational Health Sciences Institute (EOHSI)
UMDNJ – R.W. Johnson Medical School and Rutgers University
170 Fretlinghuysen Road, Piscataway, NJ 08854
Overview

• Introduction
  • 1999-2000 NHANES VOC data set
• Approaches and Results
  • Classification and Regression Tree (CART)
  • Bootstrap Aggregating (Bagging)
  • Importance measure
• Conclusions & Future Work
Introduction

- Objective: characterize key predictor variables of personal exposure to VOCs
- 1999-2000 National Health And Nutrition Examination Survey (NHANES) VOC Project Data Set
  - Number of participants: 659 adults (20-59 yrs)
  - Response variables: personal air concentrations of ten VOCs (BTEX, MTBE, TCE, PERC, Chloroform, 1,4-dichlorobenzene)
  - Predictor variables: demographic, socioeconomic, time-activity pattern
- Challenges and Limitations of the 1999-2000 NHANES VOC Data
  - Disparate data types (i.e. continuous and categorical)
  - Missing values
  - Non-detects
  - Outliers
Classification And Regression Tree (CART)

- CART models split the space of response variable values into homogenous nodes as a function of predictor variables
- CART models can handle missing values, interactions, and skewed distributions of predictor variables
- Over-fitting can be resolved by K-fold Cross-Validation
CART Analysis (Ethylbenzene)

• Response variable: personal air concentration of Ethylbenzene (ug/m³)

• VTQ200I: did you breathe fumes from or use paint thinner, brush cleaner, or furniture stripper (1=Yes, 2=No)

• VTQ110: time spent indoors at work/school
CART Analysis (o-Xylene)

- Response variable: personal air concentration of o-Xylene (ug/m³)

- VTQ090: time spent indoors at home

- RIDRETH1: race and ethnicity
  - 1 = Mexican American
  - 2 = other Hispanic
  - 3 and above = non-Hispanic
Bootstrap Aggregating (Bagging)

- Decrease the variance of model predictions by taking ensemble output of multiple instances of the base models
- Create K replicates of the training dataset and fit K tree models
- Average the predictions of the K models
Bagging Analysis (Ethylbenzene)

- Response variable: personal air concentration of Ethylbenzene (ug/m3)
- VTQ2001: did you breathe fumes from or use paint thinner, brush cleaner, or furniture stripper (1=Yes, 2=No)
- VTQ110: time spent indoors at work/school
- VTQ090: time spent indoors at home
- RIDAGEYR: age of the subject
Bagging Analysis (o-Xylene)

• Response variable: personal air concentration of Ethylbenzene (ug/m3)

• VTQ110: time spent indoors at work/school

• VTQ090: time spent indoors at home
**Concept of “Importance Measure”**

- For each predictor variable, count the number of times appearing on each of the $K$ bagging trees.
- Provide a quantitative scale for the significance of a predictor contributing to the predictive performance on the response variable.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Meaning</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTQ110</td>
<td>Time spent outdoors at work/school</td>
<td>295</td>
</tr>
<tr>
<td>INDMPIR</td>
<td>Poverty income ratio</td>
<td>69</td>
</tr>
<tr>
<td>VTQ090</td>
<td>Time spent at home</td>
<td>64</td>
</tr>
<tr>
<td>VTQ200I</td>
<td>Breathe fumes from paint thinner, brush cleaner, or furniture stripper</td>
<td>72</td>
</tr>
<tr>
<td>VTQ090</td>
<td>Time spent at home</td>
<td>70</td>
</tr>
</tbody>
</table>

Response variable: personal air conc of Ethylbenzene, $K=500$, $p=2/3$

Response variable: personal air conc of o-Xylene, $K=500$, $p=2/3$
Conclusions & Future Work

• The CART approach provides a promising way to treat the complicated data set such as the 1999-2000 NHANES VOC data
• The decision-tree structure of the CART analysis results can facilitate the interpretation of analysis outcomes
• The “Bagging” algorithm in combination with the CART approach can resolve the issue of model instability
• The importance measures resemble Bayesian posterior probabilities of predictor variables for identifying key contributors to response variables
• Future work will expand the bagging/importance measure approach to more chemicals
Acknowledgements

- U.S. EPA NERL (National Exposure Research Laboratory) through the cooperative agreement # EPAR-827033
- Mickey Leland National Urban Air Toxics Research Center