

Environmental, dietary, demographic, and activity variables associated with biomarkers of exposure for benzene and lead

A. ROY,^a P.G. GEORGOPOULOS,^a M. OUYANG,^{a,b} N. FREEMAN^a AND P.J. LIOY^a

^aExposure Measurement and Assessment Division, Environmental and Occupational Health Sciences Institute (EOHSI), 170 Frelinghuysen Road, Piscataway, New Jersey 08854, USA

^bInformatics Institute, University of Medicine and Dentistry of NJ, Piscataway, New Jersey 08854, USA

Classification and regression tree methods represent a potentially powerful means of identifying patterns in exposure data that may otherwise be overlooked. Here, regression tree models are developed to identify associations between blood concentrations of benzene and lead and over 300 variables of disparate type (numerical and categorical), often with observations that are missing or below the quantitation limit. Benzene and lead are selected from among all the environmental agents measured in the NHEXAS Region V study because they are ubiquitous, and they serve as paradigms for volatile organic compounds (VOCs) and heavy metals, two classes of environmental agents that have very different properties. Two sets of regression models were developed. In the first set, only environmental and dietary measurements were employed as predictor variables, while in the second set these were supplemented with demographic and time-activity data. In both sets of regression models, the predictor variables were regressed on the blood concentrations of the environmental agents. Jack-knife cross-validation was employed to detect overfitting of the models to the data. Blood concentrations of benzene were found to be associated with: (a) indoor air concentrations of benzene; (b) the duration of time spent indoors with someone who was smoking; and (c) the number of cigarettes smoked by the subject. All these associations suggest that tobacco smoke is a major source of exposure to benzene. Blood concentrations of lead were found to be associated with: (a) house dust concentrations of lead; (b) the duration of time spent working in a closed workshop; and (c) the year in which the subject moved into the residence. An unexpected finding was that the regression trees identified time-activity data as better predictors of the blood concentrations than the measurements in environmental and dietary media.

Journal of Exposure Analysis and Environmental Epidemiology (2003) 13, 417–426. doi:10.1038/sj.jea.7500296

Keywords: classification and regression trees, national human exposure assessment survey, benzene, lead, exposure, residential environment.

Introduction

The data collected through the National Human Exposure Assessment Survey (NHEXAS) Region V program (to be referred to as NHEXAS-V in this work) contain comprehensive information on the potential determinants of exposure to a range of heavy metals, volatile organic compounds, particulate matter, and pesticides. The large quantity of such multifaceted information presents both challenges and opportunities for scientists in the area of exposure analysis. Similar challenges are raised by other large-scale field studies that have also produced extensive data sets linking exposure and dose to environmental

variables. For example, the German nationwide environmental studies (Seifert et al., 2000a, b) have, over the past two decades, provided data on all of the above groups of contaminants (except particulate matter) for thousands of subjects. Other studies have focused on target compounds: the TEAM Study (Özkaynak et al., 1996) focused on volatile organic compounds (VOCs) (32 target VOCs in personal, indoor, and outdoor air, and drinking water for about 800 persons in eight cities, with breath measurements for every person); the Toronto study (Pellizzari et al., 1999a) focused on fine particles (180 persons with PM10 measurements, 750 with PM2.5 measurements); and the studies by Buckley and Camann (Camann et al., 1997) considered a number of pesticides in house dust of a large number of participants. A variety of flexible analytical tools is needed in order to exploit these comprehensive data sets fully and derive insights on the factors affecting the exposure processes. Much useful information has already been extracted from the NHEXAS-V data by focusing on the statistics of variables of interest, and examining the relationship between these variables and one or more relevant factors (Clayton et al., 1999; Pellizzari et al., 1999b; Thomas et al., 1999).

1. Address all correspondence to: Dr. Panos G. Georgopoulos, EOHSI - Room 308, 170 Frelinghuysen Road, Piscataway, NJ 08854, USA. Tel.: +1-732-445-0159. Fax: +1-732-445-0915.

E-mail: panosg@fidelio.rutgers.edu

Environmental and Occupational Health Sciences Institute (EOHSI) is a Joint Program sponsored by UMDNJ- Robert Wood Johnson Medical School and Rutgers University.

Received 30 May 2002; accepted 15 May 2003

A multivariate analysis examining the relationship between environmental levels, some household characteristics, and activity data, has also been conducted by examining subsets of the NHEXAS-V data (Bonnano et al., 2001). The next logical step is to attempt "mining" the NHEXAS-V data in its entirety via exploratory data analysis (EDA) methods.

Several characteristics of the NHEXAS-V data need to be recognized prior to conducting exploratory data analysis. First, the NHEXAS-V data set is comprised of a large number of variables, which may be either continuous or categorical. Second, many of the observations of the measured variables are below the analytical quantitation limit. Both these characteristics of the NHEXAS-V data set make it difficult to analyze using conventional techniques. Third, values of one or more of the large number of variables measured or surveyed may not be available for all the subjects. Therefore, it is difficult to perform exploratory data analysis using all the variables in the data set simultaneously, as the number of observations for which values are available for all the variables is generally a small fraction of the total number of observations.

In this study, Classification and Regression Tree (CART) methods were employed to perform exploratory data analysis on a large number of NHEXAS variables that are associated with benzene and lead exposures. CART methods were chosen to analyze the NHEXAS-V data set, as these methods can be used to analyze simultaneously all the variables in the data set, even though these variables are of disparate type, and have missing values, as well as observations below the quantitation limit. The variables used in this analysis were obtained from the NHEXAS Phase I field study conducted in US EPA Region V (Pellizzari et al., 1995). Benzene and lead were selected from among all the other environmental agents measured in the NHEXAS-V study because they are ubiquitous, and they serve as paradigms for VOCs and heavy metals, respectively, two classes of environmental agents that have very different properties.

CART methods have several features that can be used to advantage in describing sparse data sets with large numbers of variables of disparate types, and so they have been used in a variety of applications in several fields, including environmental, ecological, and health sciences (e.g., Breiman et al., 1984; Bouskila et al., 1998; DeFries et al., 1998; Eisenberg and McKone, 1998; Francois et al., 1998; Friedl and Brodley, 1997; Godefroy et al., 1998; Legendijk et al., 1998; Rakowski and Clark, 1998; Salzberg et al., 1998; Thiele et al., 1998; Wilson et al., 1998; Rejwan et al., 1999). Attractive features of the CART methods are: (a) They do not require the variables to be of the same type. Classification tree models in which the response variable is categorical, and regression tree models in which the response variable is continuous, can both be formulated using mixtures of continuous and categorical predictor variables. Categorical variables are permitted to have more than two levels, and are

not restricted to just logistic variables. (b) CART models are invariant to monotone transformations of the predictor variables. Therefore, CART models are not adversely influenced by predictor variables with skewed distributions. Furthermore, predictor variables having nondetectable values can be accommodated in the analysis as a consequence of this attribute of CART methods. (c) Categorical predictor variables with missing values can also be accommodated by simply treating the missing value as a level of the categorical variable. Continuous predictor variables with missing values are handled similarly after converting the variable to a categorical variable by binning the available values into an arbitrary number of ranges. This enables all the available data from an individual subject to be used in constructing the model, even though one or more of the predictor variable values are missing for the subject. (d) CART methods can handle interactions between predictor variables, without the *a priori* specification of the form of the interaction. Moreover, interactions between predictor variables in CART models are not restricted to the multiplicative form commonly employed in conventional multivariate analyses.

Methods

CART models are comprised of a collection of rules that partition the space of response variable values as a function of the predictor variables. The rules are constructed by a recursive partitioning procedure using a "training data set" containing values of the response and predictor variables. These rules can subsequently be used to predict values of the response variable, given values of the predictor variables.

Categorical predictor variables having missing values are handled by creating a new level for the categorical variable to hold all the missing values. Continuous predictor variables having missing values are first converted to categorical variables by dividing available observations into quartiles, and then adding a level to hold the missing values. These variables therefore have five levels: a level for each of the four quartiles, and a level for the missing values. One consequence of converting continuous variables to categorical variables with a level for missing values is the loss of order among the levels of the variable. Although this represents a loss of information, it can be used to advantage, as the orderings of these variables in the binary splits of a tree model provide a check on the validity of the model.

The response variables investigated in this study were concentrations of benzene and lead in blood, both of which are continuous. The regression tree models describing these variables were constructed using the one-step look-ahead binary partitioning algorithm implemented in the S-PLUS software package (Chambers and Hastie, 1992), in which the available observations of the response variable are successively partitioned by splitting existing partitions into two

groups (or partitions), starting with the entire set of observations. The rule for splitting an existing partition is prescribed by only one of the predictor variables, and the split that is optimal with respect to reduction of deviance in the response variable is selected. The deviance for partition j is defined as

$$D_j = \sum_{i=1}^{n_j} (y_{i,j} - \mu_j)^2$$

where n_j is the number of observations in partition j , $y_{i,j}$ are the observations in partition j , and μ_j is the mean of these observations. The deviance for the entire tree is calculated by summing the deviances of each partition. The size of the tree model is given by the number of terminal nodes in the tree. The size of a tree can be specified either directly or indirectly by specifying a threshold deviance or number of observations in a terminal node.

A major drawback of exploratory data analysis with CART is that the model produced by the tree growing algorithm may contain spurious associations that are merely coincidental. In other words, it is possible that the data are overfitted by the tree model, in which case the model would not be appropriate to describe another random sample from the population. This drawback was addressed by performing jack-knife cross-validation of the regression tree models to identify an optimal size for the regression trees.

Jack-knife cross-validation was performed by constructing trees of a given size using all but one observation in turn, and then calculating the average deviance of the trees of that size, computed after including the observation that was excluded in the construction of the tree. The optimal trees size was determined from the results of the cross-validation by identifying the tree size corresponding to the minimum value of the average deviance. The tree model was then constructed by specifying this tree size.

Two regression tree models were constructed for each of the environmental agents investigated in this study; benzene and lead. For each of these environmental agents, tree models were constructed to describe blood concentrations using predictor variables comprised of: (stage 1) environmental, and dietary measurements of the environmental

agent, and (stage 2) demographic variables obtained from: (a) NHEXAS-V Descriptive, Household Baseline, and Personal Baseline questionnaires, (b) time-activity data obtained or derived from Time-Activity Questionnaires, and (c) the environmental, and dietary measurements of the environmental agent that were used to develop the regression tree model in item (1) above.

Results

Summary statistics of the benzene and lead blood concentrations, and of the measurements in environmental and dietary media corresponding to these concentrations, are given in Tables 1 and 2, respectively. The summary statistics pertain only to the subset of the NHEXAS-V data for which corresponding measurements were available for benzene and lead blood concentrations. The numbers of available measurements used to calculate the 25th, 50th, and 75th percentiles for each environmental and dietary media variable, and the percentages of these measurements that were quantifiable are also presented. The summary statistics show that 88% of the 143 available observations of benzene concentrations in blood, and 97% of the 165 available observations of lead concentrations in blood, are above the quantitation limit. The numbers of available measurements in environmental and dietary media corresponding to these blood concentrations range from 36 to 143 for benzene, and 10 to 164 for lead (excluding the media for which no measurements were available). The percentages of the available environmental and dietary measurements that were above the quantitation limit range from 5.6% to 100% for benzene, and 24% to 100% for lead. The means and variances of the measurements are not reported, as these would be affected by default values assigned to observations below the quantitation limit in the NHEXAS-V data sets.

The Pearson and Spearman correlations between blood concentrations and environmental and dietary measurements are also reported in Tables 1 and 2 for benzene and lead, respectively, and a scatter-histogram plot of benzene observations is presented in Figure 1. The correlation

Table 1. Summary statistics of benzene concentrations in blood and the corresponding measurements in environmental and dietary media (observations include estimates for non-detects).

| Variable | Description | No. obs. | % Quantifiable | First quartile | Median | Third quartile | Pearson's correlation | Spearman's correlation |
|----------|--|----------|----------------|----------------|----------|----------------|-----------------------|------------------------|
| MEAS210 | Blood conc. ($\mu\text{g}/\text{l}$) | 143 | 88.1 | 8.60E-02 | 1.50E-01 | 3.30E-01 | 1.00 | 1.00 |
| MEAS160 | Personal air conc. ($\mu\text{g}/\text{m}^3$) | 140 | 100.0 | 3.55E+00 | 5.59E+00 | 7.89E+00 | 0.58 | 0.31 |
| MEAS170 | Non-work personal air conc. ($\mu\text{g}/\text{m}^3$) | 36 | 100.0 | 5.06E+00 | 8.27E+00 | 1.02E+01 | 0.09 | 0.27 |
| MEAS175 | Work personal air conc. ($\mu\text{g}/\text{m}^3$) | 36 | 52.8 | 0.00E+00 | 2.35E-01 | 2.18E+00 | 0.03 | -0.19 |
| MEAS180 | Indoor air conc. ($\mu\text{g}/\text{m}^3$) | 143 | 99.3 | 3.18E+00 | 4.69E+00 | 7.03E+00 | 0.53 | 0.30 |
| MEAS190 | Outdoor air conc. ($\mu\text{g}/\text{m}^3$) | 59 | 100.0 | 2.16E+00 | 3.00E+00 | 4.75E+00 | 0.66 | -0.14 |
| MEAS200 | Drinking water conc. ($\mu\text{g}/\text{l}$) | 143 | 5.6 | 0.00E+00 | 4.99E-03 | 1.27E-02 | 0.08 | 0.04 |

Table 2. Summary statistics of lead concentrations in blood, and the corresponding lead measurements in environmental and dietary media (observations include estimates for non-detects).

| Variable | Description | No. obs. | % Quantifiable | First quartile | Median | Third quartile | Pearson's correlation | Spearman's correlation |
|----------|---|----------|----------------|----------------|----------|----------------|-----------------------|------------------------|
| MEAS140 | Blood ($\mu\text{g/l}$) | 165 | 97.0 | 1.00E+00 | 1.70E+00 | 2.80E+00 | 1.00 | 1.00 |
| MEAS010 | Personal air conc. (ng/m^3) | 126 | 80.2 | 7.32E+00 | 1.30E+01 | 2.59E+01 | 0.15 | 0.17 |
| MEAS020 | Indoor air conc. (ng/m^3) | 153 | 49.7 | 3.69E+00 | 6.30E+00 | 1.14E+01 | -0.01 | 0.17 |
| MEAS021 | Indoor air conc. PM_{10} (ng/m^3) | 25 | 24.0 | 4.17E+00 | 6.50E+00 | 1.22E+01 | 0.02 | 0.19 |
| MEAS030 | Outdoor air conc. (ng/m^3) | 66 | 72.7 | 5.99E+00 | 8.63E+00 | 1.23E+01 | -0.04 | -0.07 |
| MEAS031 | Outdoor air conc. PM_{10} (ng/cm^2) | 21 | 23.8 | 4.86E+00 | 7.51E+00 | 1.03E+01 | -0.08 | -0.18 |
| MEAS050 | LWW surface dust loading (ng/cm^2) | 164 | 91.5 | 1.68E+00 | 5.94E+00 | 1.97E+01 | 0.01 | 0.25 |
| MEAS052 | LWW surface dust conc. ($\mu\text{g/g}$) | 163 | 91.4 | 6.83E+01 | 1.29E+02 | 3.03E+02 | 0.26 | 0.17 |
| MEAS053 | WWT surface dust loading (ng/cm^2) | 45 | 91.1 | 5.30E+00 | 1.16E+01 | 2.54E+01 | 0.19 | 0.10 |
| MEAS055 | LWW window sill dust loading (ng/cm^2) | 160 | 94.4 | 4.17E+00 | 1.28E+01 | 6.69E+01 | 0.10 | 0.15 |
| MEAS057 | LWW window sill dust conc. ($\mu\text{g/g}$) | 160 | 94.4 | 9.05E+01 | 1.88E+02 | 5.18E+02 | 0.22 | 0.10 |
| MEAS058 | WWT window sill dust loading (ng/cm^2) | 44 | 97.7 | 7.57E+00 | 3.04E+01 | 7.90E+01 | -0.12 | -0.04 |
| MEAS070 | Entranceway soil conc. ($\mu\text{g/g}$) | 41 | 87.8 | 2.61E+01 | 7.51E+01 | 3.17E+02 | 0.23 | 0.26 |
| MEAS080 | Yard soil conc. ($\mu\text{g/g}$) | 41 | 97.6 | 1.97E+01 | 4.19E+01 | 3.01E+02 | 0.45 | 0.32 |
| MEAS090 | Standing tap water conc. ($\mu\text{g/l}$) | 163 | 100.0 | 7.44E-01 | 2.24E+00 | 4.54E+00 | 0.02 | 0.09 |
| MEAS100 | Flushed tap water conc. ($\mu\text{g/l}$) | 164 | 83.5 | 1.36E-01 | 3.03E-01 | 6.42E-01 | 0.19 | 0.19 |
| MEAS110 | Bottled water conc. ($\mu\text{g/l}$) | 10 | 30.0 | 2.68E-02 | 5.00E-02 | 1.02E-01 | -0.30 | -0.15 |
| MEAS111 | Day 1 solid food conc. ($\mu\text{g/kg}$) | 23 | 100.0 | 4.05E+00 | 5.90E+00 | 7.25E+00 | -0.02 | 0.03 |
| MEAS112 | Day 1 beverages conc. ($\mu\text{g/kg}$) | 23 | 95.7 | 5.65E-01 | 9.10E-01 | 1.45E+00 | 0.01 | 0.03 |
| MEAS113 | Day 2 solid food conc. ($\mu\text{g/kg}$) | 23 | 95.7 | 4.65E+00 | 6.10E+00 | 8.35E+00 | -0.19 | -0.61 |
| MEAS114 | Day 2 beverages conc. ($\mu\text{g/kg}$) | 23 | 87.0 | 5.05E-01 | 9.10E-01 | 1.25E+00 | 0.04 | 0.10 |
| MEAS115 | Day 3 solid food conc. ($\mu\text{g/kg}$) | 23 | 100.0 | 5.65E+00 | 6.80E+00 | 9.50E+00 | 0.16 | 0.21 |
| MEAS116 | Day 3 beverages conc. ($\mu\text{g/kg}$) | 23 | 82.6 | 4.20E-01 | 7.60E-01 | 1.68E+00 | 0.02 | 0.09 |
| MEAS117 | Day 4 solid food conc. ($\mu\text{g/kg}$) | 23 | 95.7 | 4.10E+00 | 5.70E+00 | 8.55E+00 | -0.05 | -0.25 |
| MEAS118 | Day 4 beverages conc. ($\mu\text{g/kg}$) | 23 | 73.9 | 3.40E-01 | 6.10E-01 | 1.26E+00 | 0.06 | 0.18 |
| MEAS120 | Solid food conc. ($\mu\text{g/kg}$) | 130 | 100.0 | 5.90E+00 | 7.08E+00 | 9.58E+00 | 0.12 | 0.08 |
| MEAS121 | Day 1 food + beverage conc. ($\mu\text{g/kg}$) | 23 | 100.0 | 2.14E+00 | 2.79E+00 | 3.34E+00 | -0.01 | 0.17 |
| MEAS122 | Day 2 food + beverage conc. ($\mu\text{g/kg}$) | 23 | 100.0 | 2.27E+00 | 2.66E+00 | 3.02E+00 | -0.17 | -0.35 |
| MEAS123 | Day 3 food + beverage conc. ($\mu\text{g/kg}$) | 23 | 100.0 | 2.50E+00 | 2.79E+00 | 3.73E+00 | 0.08 | 0.04 |
| MEAS124 | Day 4 food + beverage conc. ($\mu\text{g/kg}$) | 23 | 100.0 | 1.93E+00 | 2.77E+00 | 3.06E+00 | -0.06 | -0.19 |
| MEAS130 | Beverages conc. ($\mu\text{g/kg}$) | 130 | 92.3 | 6.73E-01 | 1.10E+00 | 1.89E+00 | 0.18 | 0.06 |
| MEAS132 | Food + beverages conc. ($\mu\text{g/kg}$) | 127 | 100.0 | 2.27E+00 | 3.18E+00 | 4.49E+00 | 0.19 | 0.09 |
| MEAS134 | Food intake ($\mu\text{g/day}$) | 130 | 100.0 | 3.36E+00 | 4.94E+00 | 7.30E+00 | -0.01 | 0.00 |
| MEAS136 | Beverage intake ($\mu\text{g/day}$) | 130 | 92.3 | 9.55E-01 | 1.69E+00 | 3.23E+00 | 0.16 | 0.03 |
| MEAS138 | Food + bev. intake ($\mu\text{g/day}$) | 127 | 100.0 | 4.87E+00 | 7.31E+00 | 1.20E+01 | 0.13 | 0.08 |

coefficients are calculated based only on the observations for which corresponding benzene and lead blood concentrations measurements were available (the numbers of available observations for each variable are given in Tables 1 and 2). Although the Pearson correlations between benzene blood concentrations and measurements in all six media are positive, the Spearman Rank Correlations for two of these media (*Work Personal Air Concentration* and *Outdoor Air Concentration*) are slightly negative. The high Pearson correlation for *Outdoor Air Concentration* appears to be driven by a single observation (see Figure 1). Only the correlations for *Personal Air Concentration* and *Indoor Air Concentration* appear to be notable. The Pearson and Spearman correlation coefficients between lead blood concentration and all other environmental and dietary media were both highest for *Yard Soil Concentration* (0.45 and 0.32, respectively). Although negative correlations were observed for some variables, these were either based on a small number

of observations, or the Pearson and Spearman correlations were not consistent. As mentioned earlier, the regression tree models for benzene and lead concentrations in blood were developed in two stages: the first stage models were based only on the environmental and dietary measurements summarized in Tables 1 and 2, and the second stage models also included demographic and time-activity data. The demographic data were obtained from answers to questions compiled from the NHEXAS-V Descriptive, Household Baseline, and Personal Baseline questionnaires and have been organized in a table, which is posted online as a supplement to this article (<http://www.ccl.rutgers.edu/NHEXAS-V/index.html>). The time-activity variables used in developing the second-stage regression tree models are listed in Table 3.

Regression tree models that do not overfit the data were identified using jack-knife cross-validation. The results of the jack-knife cross-validation of regression tree models are

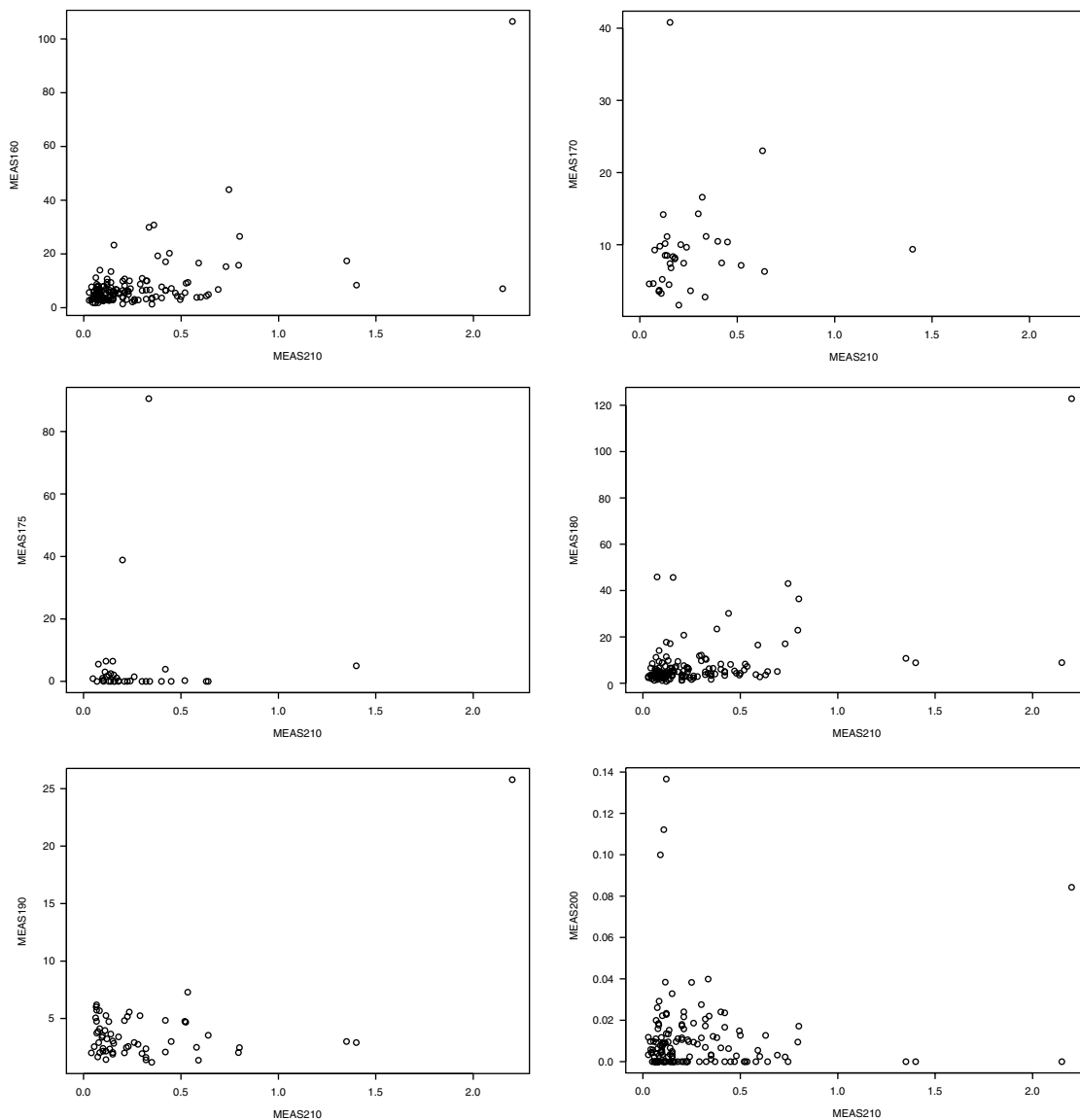


Figure 1. Scatter plots of benzene concentrations in environmental and biological media. (Variables are defined in Table 1.)

presented by plotting average deviance as a function of tree size (number of terminal nodes in the tree). The average deviance of trees of a given size was calculated from the deviance of the N trees formulated using training data sets comprised of all but one of the N available observations in turn. The jack-knife cross-validation procedure was implemented by calculating the deviance of each of the N trees of a given size using all the N available observations, although each tree was formulated using only $N-1$ observations. The optimal size for a tree describing the data is suggested by the size corresponding to trees displaying the lowest average deviance.

The results of the jack-knife cross-validation of regression tree models for benzene concentration in blood as a function of environmental and dietary measurements of benzene is

shown in Figure 2a. These tree models are based on measurements of benzene in the six environmental and dietary media given in Table 1, that correspond to the available measurements of benzene in blood. The minimum average deviance in Figure 2a remains almost constant for trees of sizes 1 and 2, and increases thereafter, suggesting that the variance in benzene blood concentrations is not well explained by environmental and dietary measurements of benzene, and that a tree having at most one binary split can be justified to explain the variance in benzene blood concentrations.

A dendrogram representing the two-terminal node regression tree model that best describes benzene blood concentrations, based on all of the 143 available environmental and dietary measurements of benzene, is shown in

Table 3. Time-activity variables used in constructing tree models.

| Variable | Description |
|----------|--|
| TE1 | Av. h/day inside at home |
| TE2 | Av. h/day inside at work/school |
| TE3 | Av. h/day inside elsewhere |
| TE4 | Av. h/day outside at home |
| TE5 | Av. h/day outside at work/school |
| TE6 | Av. h/day outside elsewhere |
| TE7 | Av. h/day in transit |
| TE123 | Av. h/day indoors |
| TE456 | Av. h/day outdoors |
| TA1 | No. days/week pumped gas |
| TA2 | No. days/week gasoline on skin |
| TA3 | No. days/week in enclosed garage with car |
| TA4 | No. days/week with yard dirt/soil on skin |
| TA5 | No. days/week grass/leaves on skin |
| TA6 | No. days/week cleaned fireplace/wood stove |
| TA7 | No. days/week started/tended fire |
| TA8 | No. days/week used outdoor grill/burn debris |
| TA9 | No. days/week tobacco smoked in home |
| TA10 | No. days/week took shower |
| TA11 | No. days/week took bath |
| TA12 | Av. no. glasses of water/week |
| TA13 | Av. number cigarettes smoked/week |
| TA14 | Av. number cigars/pipefuls smoked/week |
| TA15 | Av. no. times used smokeless tobacco/week |
| TA16 | Av. no. times washed hands/week |
| TA17 | Av. min/week traveled on roadways/highways |
| TA18 | Av. min/week indoors with smoker |
| TA19 | Av. min/week in vehicle with smoker |
| TA20 | Av. min/week swam in pool |
| TA21 | Av. min/week used cleaning supplies |
| TA22 | Av. min/week sat/lay on carpet/rugs |
| TA23 | Av. min/week in enclosed workshop |
| TA24 | Av. min/week doors and windows left open |
| TA25 | Av. min/week performed vigorous exercise |
| TA26 | Av. min/week performed moderate exercise |

Figure 2b. The single binary split in the model is based on *MEAS180* (*Indoor Air Concentration*), and the condition shown in the dendrogram ($MEAS180 < 7.87 \mu\text{g}/\text{m}^3$) defines the left split.

The average value of benzene concentration in blood in the two terminal nodes is shown under the nodes, along with histograms of the quartile distribution of values in each node. The dendrogram shows that all the benzene blood concentrations in Node 2 are in the highest quartile of the overall distribution, but that most of the observations fall under Node 1, including many observations in the highest quartile. Although it is possible to explain the distribution of values in Node 1 by trees of larger size in which Node 1 is partitioned further, these larger tree models could not be justified given the results of the jack-knife cross-validation. Nonetheless, it is of interest to note that all the splits in trees having up to five terminal nodes were based on *Indoor Air Concentration* measurements.

The results of the jack-knife cross-validation of second-stage regression tree models that describe benzene concentra-

tions in blood in terms of environmental and dietary media measurements of benzene, as well as demographic and time-activity data are shown in Figure 2c. The cross-validation suggests that benzene concentrations in blood in this data set are optimally described by a tree with four terminal nodes, and that trees of this size account for an average of approximately 50% of the overall deviance in the blood concentration. The best regression tree for benzene concentration in blood with four terminal nodes, generated using the 143 available measured values of benzene blood concentration and the corresponding values of 319 predictor variables, is shown in Figure 2d. The lengths of the vertical lines under a split are proportional to the reduction of overall deviance produced by the split. The variables (and conditions) defining the left splits at the three internal nodes of the model are (in order of importance): (1) the average number of min/day spent indoors with someone who was smoking ($TA18 < 914$); (2) a categorical variable specifying the range of years in which apartment/house was built ($AB23:fgi$, where the *f*, *g*, and *i* categories represent 1950–1959, 1940–1949, and “missing”, respectively); and (3) the average number of cigarettes/day smoked by the subject ($TA13 < 3.4$). The right split corresponding to the *AB23* variable is comprised of the ranges 1970–1979, 1960–1969, and 1939 or earlier; there were no observations in the ranges 1980–1989, and 1990 and later.

The results of the jack-knife cross-validation of regression tree models for lead concentration in blood, having predictor variables comprised of lead environmental and dietary media measurements, are shown in Figure 3a. These tree models were based on measurements of lead in the 38 environmental and dietary media given in Table 2. The results of the cross-validation suggest that the deviance in blood lead concentration is optimally described by a regression tree having two terminal nodes, as the average deviance is minimum for trees of this size.

The dendrogram shown in Figure 3b represents the two-terminal node regression tree model that best describes lead concentrations in blood, based on all of the 165 available environmental and dietary measurements of lead. The single binary split in the model is based on *MEAS050* (*House Dust Concentration*), which was treated as a categorical variable, as observations for this variable are not available for all 165 observations of blood lead concentration. The available observations were categorized by placing them in quartiles (first, second, third, and fourth quartiles being represented by *a*, *b*, *c*, and *d*), and the missing observations were placed in a separate category (represented by *e*). The left split in the dendrogram is defined by the condition above the split ($MEAS050:abce$), which implies that Node 2 contains all observations for which *MEAS050* values were available and were in the highest quartile for these observations. The average blood lead concentrations under Nodes 1 and 2 are 1.962 and 3.195 $\mu\text{g}/\text{dl}$ respectively. The quartile histograms under the

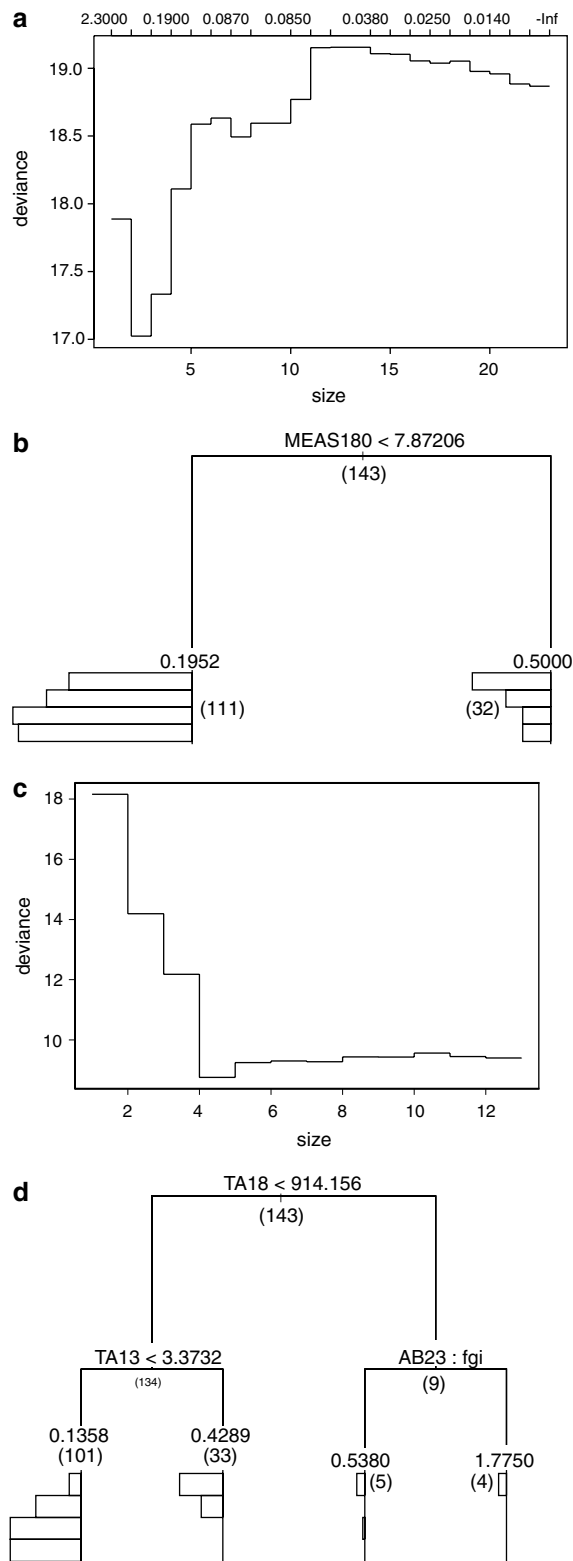
terminal nodes show that the fraction of each quartile in Node 2 increases monotonically with respect to the quartile, with an especially large increase in the fraction in the fourth quartile in

comparison to the third quartile. It is relevant to note that the split condition is consistent with the relationship between the categories of *MEAS050* ($a < b < c < d$), even though the relationship between the categories of the *MEAS050* variables was not used by the tree building algorithm (category *e* of *MEAS050* is comprised of missing values).

The results of the jack-knife cross-validation of second-stage regression tree models that describe blood lead concentrations in terms of environmental and dietary media measurements of lead, and of demographic and time-activity data, are shown in Figure 3c. The cross-validation indicates that the deviance in blood lead concentration is optimally described by a regression tree having three terminal nodes, and that trees of this size explain approximately 8% of the overall deviance on average. The best regression tree model having three terminal nodes is shown in Figure 3d. The variables (conditions) defining the splits at the two internal nodes are: (1) the average min/day spent in a closed workshop ($TA23 < 44$) and (2) a categorical variable specifying the period (range of years) in which the subject moved into the apartment/house ($AB24:abc$, where the *a*, *b*, and *c* categories represent 1990 or later, 1980–1989, and 1970–1979, respectively).

Discussion

The application of tree-based methods to explore large exposure databases, such as those generated under the



NHEXAS program, was demonstrated using data associated with benzene and lead exposures from the US EPA Region V

NHEXAS Phase I data set. Benzene and lead were selected from among all the environmental agents measured in the study because they are ubiquitous, and they serve as paradigms for VOCs and heavy metals, respectively, two classes of environmental agents that have very different properties. The response variable chosen to be described by the tree models for both benzene and lead was blood concentration, the only biomarker measured for these agents. Biomarkers are an attractive measure of exposure, as they have a causal link to internal dose. However, biomarker data are often difficult to obtain, and the relationship between biomarkers of exposure and dose is usually not straightforward (Weisel et al., 1996; Roy and Georgopoulos, 1998), as is evident by considering the factors that affect the blood concentrations of benzene and lead. Exposure to benzene generally occurs via inhalation, which has an immediate impact on blood concentrations, whereas lead exposure generally occurs via ingestion, which has a delayed effect on blood concentration, as the ingested lead is absorbed into systemic circulation over a period of time. Consequently, benzene blood concentrations are highly dependent upon the time that elapses between the end of exposure and sampling time, with concentration levels dropping rapidly after the exposure ends, whereas the within subject variability in blood lead concentrations is relatively lower. The terminal half-life of benzene is much shorter than that of lead, as the terminal half-life for benzene depends on the rates of release from fat tissue, whereas the terminal half life for lead depends upon

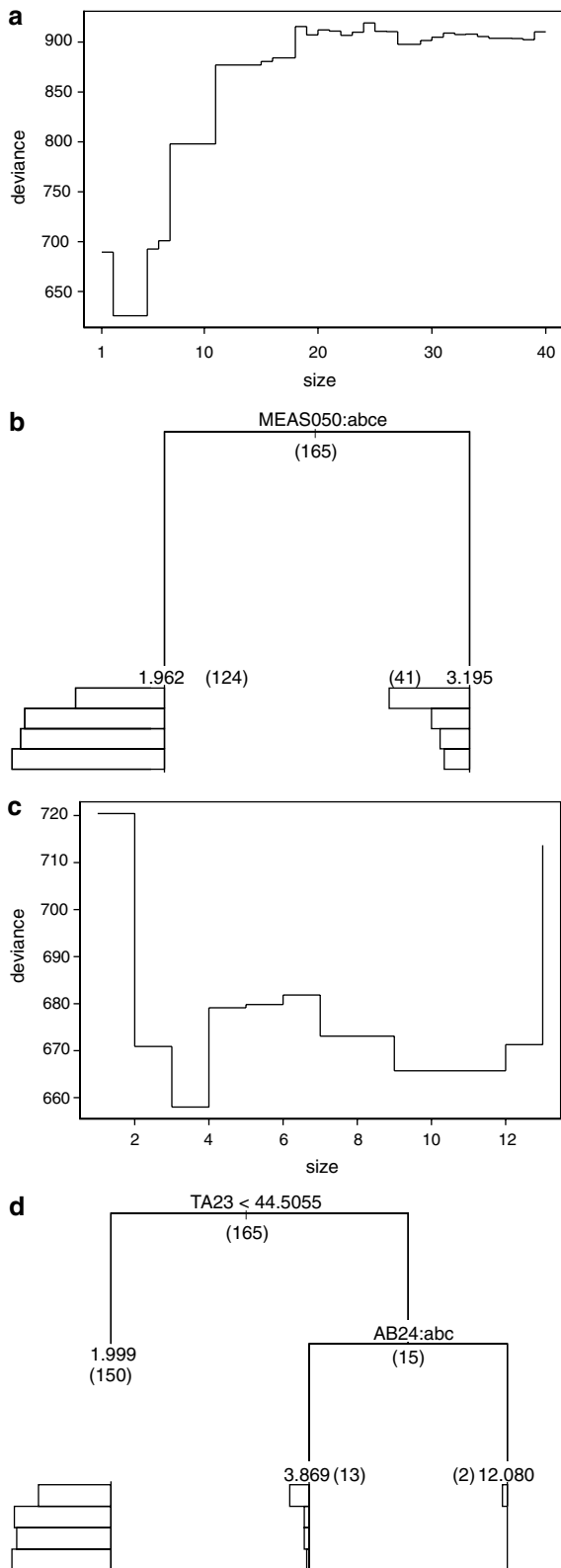


Figure 3. (a) Average deviance of lead concentration in blood ($\mu\text{g}/\text{dl}$)² computed by performing cross-validation on regression trees of varying size that are constructed using measurements of lead in environmental and dietary media. The number of terminal nodes (size) of the optimal tree model is identified to be 2, based on the minimum value of the average deviance. (b) Dendrogram of the optimal regression tree describing lead concentration in blood, as a function of lead measurements in environmental, dietary media. The only predictor variable is MEAS050 (House Dust Concentration). The distribution of lead blood concentrations in each terminal node is shown graphically as fractions of the quartiles of the overall distribution, and the average value ($\mu\text{g}/\text{dl}$) is given under the node. The numbers in parentheses represent the number of observations at each split/node. (c) Average deviance of lead concentrations in blood ($\mu\text{g}/\text{dl}$)² computed by performing jack-knife cross-validation on regression trees of varying size. The regression trees are based on lead concentrations in environmental and dietary media, baseline questionnaires, and time-activity data. The number of terminal nodes (size) of the optimal tree model is identified to be 3, based on the minimum value of the average deviance. (d) Dendrogram of the optimal regression tree describing the concentration of lead in blood, as a function of environmental and dietary measurements of lead, as well as baseline and descriptive questionnaires and time-activity diary data. The values of the variables specifying the left branch of a split are shown above the split. The distribution of lead concentrations in each terminal node is shown graphically under the dendrogram and the average value ($\mu\text{g}/\text{dl}$) is given under the node. The predictor variables in the tree model are: (a) TA23: average min/day spent in a closed workshop; and (b) AB24: when moved into apartment/house. The numbers in parentheses represent the number of observations at each split/node.

the rates of release from bone (Wallace, 1989; Gordon et al., 1996). It is therefore of interest to ascertain whether an association can be established between these biomarkers of benzene and lead exposure and the environmental, demographic, and time-activity variables measured in NHEXAS-V.

The regression tree model for benzene in the first set suggests that the variability of benzene concentrations in blood is best explained by the indoor air concentrations of benzene. This is an unexpected result, as the correlations between blood concentration and both (a) outdoor air concentration, and (b) personal air concentration are higher than the correlation between blood concentration and indoor air concentration. The probable explanation is that correlation is a global measure of association, whereas the tree models identify local associations. It is encouraging to note that although the predictor variables in the first- and second-stage regression tree models for benzene are different, they both suggest that elevated blood concentrations of benzene are associated with the same pathway; namely, inhalation exposure of cigarette smoke. It is also worthwhile to note that, in the second-stage regression tree model for benzene, the variability in benzene blood concentration was best explained by *TA18*, a variable that can be considered to be a surrogate measure of exposure to second-hand smoke, and that *TA13*, which can be considered to be a measure of direct inhalation of cigarette smoke, was less important. The significance of the third node in the tree, which depends upon the *AB23* variable, is unclear. The split at this node suggests that of participants who experienced extended exposure to second-hand smoke indoors, those who lived in homes built after 1950 had higher benzene concentrations in blood.

The first-stage regression tree model for lead reaffirms the relationship between lead concentrations in house dust and blood, that has been noted in several studies (Lioy et al., 1998; Rhoads et al., 1999). However, as in the case of benzene, blood concentrations of lead appear to be most associated with activity data. High levels of lead concentration in blood are associated with the time spent in a closed workshop. Furthermore, of the individuals who spent more than approximately 45 min/day in a closed workshop, those who moved into their apartment after 1980 had lower blood lead concentrations on average. The significance of this last finding is not clear, although it may be related to the discontinuation of the use of leaded gasoline in the US. The fact that the splitting condition is consistent despite being a categorical variable suggests that this association is not an artifact.

In conclusion, it has been demonstrated that CART methods are a useful tool for performing exploratory data analysis on a large number of exposure-relevant variables of disparate type, many of which have values that are below the quantitation limit or are missing. Moreover, it has been shown that these methods can be useful in generating hypotheses about exposure pathways that are not obvious.

An unexpected finding is that the regression trees identified time-activity data as better predictors of the blood concentrations than the measurements in environmental and dietary media. However, the possibility that some of these associations are artifacts cannot be ruled out, despite the precautions taken to identify false associations by using cross-validation diagnostic tests. The low number of associations reported is a result of this diagnostic effort. The apparent lack of association between environmental measures and blood concentrations may result from the limitations in sampling timing and durations of the NHEXAS-V study. The analysis presented here was conducted using a limited number of observations, which may result in patterns that are coincidental, and not representative of the population as a whole. However, the number of associations identified using CART should increase as the number of observations analyzed increases, as will the confidence in the validity of these associations.

The sampling weights in the NHEXAS-V data set were not used in the analyses presented here, and therefore the results reflect relationships in the observed individuals. Therefore, although these results are suggestive of relationships in the underlying NHEXAS-V population, they cannot be readily generalized to this population.

Acknowledgments

We acknowledge the contribution of Drs. Richard Opiekun and James Quackenboss, for their help in understanding the NHEXAS-V data, and the USEPA for funding the work (NHEXAS Co-operative Agreement CR827713 and EPA University Partnership CR827033). This work was also partly funded by the NIEHS Center at EOHHSI (P30ES05022). This work has not yet been subjected to EPA peer review.

References

- Bonnano L.J., Freeman N.C.G., Greenberg M., and Lioy P.J. Multivariate analysis on levels of selected metals, particulate matter, VOC, and household characteristics and activities from Midwestern States NHEXAS. *Appl Occup Environ Hygiene* 2001; 16(9): 1-16.
- Bouskila A., Robinson M.E., Roitberg B.D., and Tenhumberg B. Life-history decisions under predation risk: importance of a game perspective. *Evol Ecol* 1998; 12: 701-715.
- Breiman L., Friedman J.H., Olshen R.H., and Stone C.J. *Classification and Regression Trees. The Wadsworth Statistics/Probability Series.* Wadsworth International Group, Belmont, CA, 1984.
- Camann D.E., Akland G.G., Buckley J.D., Bond A.E., and Mage D.T. Carpet Dust and Pesticide Exposure of Farm Children. International Society of Exposure Analysis Meeting, Research Triangle Park, NC, 1997.
- Chambers J.M., and Hastie T.J. *Statistical Models in S. Wadsworth and Brooks/Cole Advanced Books and Software.* Pacific Grove, CA, 1992.

- Clayton C.A., Pellizzari E.D., Whitmore R.W., Perritt R.L., and Quackenboss J.J. National Human Exposure Assessment Survey (NHEXAS): distributions and associations of lead, arsenic and volatile organic compounds in EPA region 5. *J Expos Anal Environ Epidemiol* 1999; 9(5): 381–392.
- DeFries R.S., Hansen M., Townshend J.R.G., and Sohlberg R. Global land cover classifications at 8 km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers. *Int J Remote Sensing* 1998; 19: 3141–3168.
- Eisenberg J.N.S., and McKone T.E. Decision tree method for the classification of chemical pollutants: incorporation of across chemical variability and within-chemical uncertainty. *Environ Sci Technol* 1998; 32: 3396–3404.
- Francois C., Rummelink M., Petein M., vanVelthoven R., Danguy A., Wespes E., Salmon I., Kiss R., and Decaestecker C. The chromatin pattern of cell nuclei is of prognostic value for renal cell carcinomas. *Anal Cell Pathol* 1998; 16: 161–175.
- Friedl M.A., and Brodley C.E. Decision tree classification of land cover from remotely sensed data. *Remote Sensing Environ* 1997; 61: 399–409.
- Godefroy O., Duhamel A., Leclerc X., SaintMichel T., Henon H., and Leys D. Brain-behaviour relationships — some models and related statistical procedures for the study of brain-damaged patients. *Brain* 1998; 121: 1545–1556.
- Gordon S.M., Callahan P.J., Wallace L.A., and Pleil J.D. Continuous Real-time Analysis of Volatile Organic Compounds to Determine Exposure-Dose Relationships. Annual Meeting of SRA/ISEA, New Orleans, LA, 1996.
- Lagendijk J.H., Mullink H., VanDiest P.J., Meijer G.A., and Meijer C.J.L.M. Tracing the origin of adenocarcinomas with unknown primary using immunohistochemistry: differential diagnosis between colonic and ovarian carcinomas as primary sites. *Hum Pathol* 1998; 29: 491–497.
- Lioy P.J., Yiin L.M., Adgate J., Weisel C., and Rhoads G.G. The effectiveness of a home cleaning intervention strategy in reducing potential dust and lead exposures. *J Expos Anal Environ Epidemiol* 1998; 8(1): 17–35.
- Özkaynak H., Xue J., Spengler J.D., Wallace L.A., Pellizzari E.D., and Jenkins P. Personal exposure to airborne particles and metals: results from the Particle TEAM Study in Riverside, CA. *J Expos Anal Environ Epidemiol* 1996; 6: 57–78.
- Pellizzari E.D., Clayton C.A., Rodes C.E., Mason R.E., Piper L.L., Fort B., Pfeifer G., and Lynam D. Particulate matter and manganese exposures in Toronto, Canada. *Atmos Environ* 1999a; 33: 721–734.
- Pellizzari E.D., Lioy P., Quackenboss J., Whitmore R., Clayton C.A., Freeman N., Waldman J., Thomas K., Rodes C., and Wilcosky T. Population-based exposure measurements in EPA region 5: a phase I field study in support of the National Human Exposure Assessment Survey. *J Expos Anal Environ Epidemiol* 1995; 5(3): 327–358.
- Pellizzari E.D., Perritt R.L., and Clayton C.A. National human exposure assessment survey (NHEXAS): exploratory survey of exposure among population subgroups in EPA Region V. *J Expos Anal Environ Epidemiol* 1999b; 9(1): 49–55.
- Rakowski W., and Clark M.A. Do groups of women aged 50 to 75 match the national average mammography rate? *Am J Prev Med* 1998; 15: 187–197.
- Rejwan C., Collins N.C., Brunner L.J., Shuter B.J., and Ridgway M.S. Tree regression analysis on the nesting habitat of smallmouth bass. *Ecology* 1999; 80: 341–348.
- Rhoads G.G., Ettinger A.S., Weisel C.P., Buckley T.J., Goldman K.D., Adgate J., and Lioy P.J. The effect of dust lead control on blood lead in toddlers: a randomized trial. *Pediatrics* 1999; 103(3): 551–555.
- Roy A., and Georgopoulos P.G. Reconstructing week-long exposures to volatile organic compounds using physiologically based pharmacokinetic models. *J Expos Anal Environ Epidemiol* 1998; 8(3): 407–422.
- Salzberg S., Delcher A.L., Fasman K.H., and Henderson J. A decision tree system for finding genes in DNA. *J Comput Biol* 1998; 5: 667–680.
- Seifert B., Becker K., Helm D., Krause C., Schulz C., and Seiwert M. The German Environmental Survey 1990/1992 (GerES II): reference concentrations of selected environmental pollutants in blood, urine, hair, house dust, drinking water, and indoor air. *J Expos Anal Environ Epidemiol* 2000a; 10: 552–565.
- Seifert B., Becker K., Hoffman K., Krause C., and Schulz C. The German Environmental Survey 1990/1992 (GerES II): a representative population study. *J Expos Anal Environ Epidemiol* 2000b; 10: 103–114.
- Thiele J., Kvasnicka H.M., Zirbes T.K., Flucke U., Niederle N., Leder L.D., Diehl V., and Fischer R. Impact of clinical and morphological variables in classification and regression tree-based survival (CART) analysis of CML with special emphasis on dynamic features. *Eur J Haematol* 1998; 60: 35–46.
- Thomas K.W., Pellizzari E.D., and Berry M.R. Population-based dietary intakes and tap water concentrations for selected elements in the EPA region V National Human Exposure Assessment Survey (NHEXAS). *J Expos Anal Environ Epidemiol* 1999; 9(5): 402–413.
- Wallace L.A. The exposure of the general population to benzene. *Cell Biol Toxicol* 1989; 5: 297–314.
- Weisel C., Yu R., Roy A., and Georgopoulos P. Biomarkers of environmental benzene exposure. *Environ Health Perspect* 1996; 104(Suppl. 6): 1141–1146.
- Wilson W.H., Zierzow R.E., and Savage A.R. Habitat selection by peatland birds in a central Maine bog: the effects of scale and year. *J Field Ornithol* 1998; 69: 540–548.