NIH Public Access
Author Manuscript
Inhal Toxicol. Author manuscript; available in PMC 2013 May 02.

Published in final edited form as:
Inhal Toxicol. 2011 August ; 23(0 2): 31–41. doi:10.3109/08958378.2010.566291.

The Toxicological Evaluation of Realistic Emissions of Source Aerosols Study: Statistical Methods

Brent A. Coull, **Gregory A. Wellenius**, **Beatriz Gonzalez-Flecha**, **Edgar Diaz**, **Petros Koutrakis**, and **John J. Godleski**
Departments of Biostatistics & Environmental Health Harvard School of Public Health 655 Huntington Avenue Boston, Massachusetts 02115, U. S. A.

Abstract

The Toxicological Evaluation of Realistic Emissions of Source Aerosols (TERESA) study involved withdrawal, aging, and atmospheric transformation of emissions of three coal-fired power plants. Toxicological evaluations were carried out in rats exposed to different emission scenarios with extensive exposure characterization. Data generated had multiple levels of resolution: exposure, scenario and constituent chemical composition. Here, we outline a multilayered approach to analyze the associations between exposure and health effects beginning with standard ANOVA models that treat exposure as a categorical variable. The model assessed differences in exposure effects across scenarios (by plant). To assess unadjusted associations between pollutant concentrations and health, univariate analyses were conducted using the difference between the response means under exposed and control conditions and a single constituent concentration as the predictor. Then, a novel multivariate analysis of exposure composition and health was used based on random forests, a recent extension of classification and regression trees that were applied to the outcome differences. For each exposure constituent, this approach yielded a nonparametric measure of the importance of that constituent in predicting differences in response on a given day, controlling for the other measured constituent concentrations in the model. Finally, an R^2 analysis compared the relative importance of exposure scenario, plant, and constituent concentrations on each outcome. Peak expiratory flow is used to demonstrate how the multiple levels of the analysis complement each other to assess constituents most strongly associated with health effects.

Keywords

additive mixed model; complex mixture; high-dimensional predictor; random forest

1. INTRODUCTION

The Toxicological Evaluation of Realistic Emissions of Source Aerosols (TERESA) study presents a unique opportunity to assess the potential health effects resulting from exposure to primary and secondary particulate emissions of coal fired power plants. The study involves withdrawal of emissions directly from the stacks of three coal-fired power plants in the US. The emissions were aged and atmospherically transformed in a mobile laboratory that simulated downwind power plant plume processing. This approach produced exposure scenarios of increasing complexity by adding atmospheric constituents, including alpha-pinene as a secondary organic aerosol (SOA) precursor, and ammonia, in a step-wise fashion. Extensive exposure characterization was carried out, including gas-phase and

bcoull@hsph.harvard.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

particulate species, and toxicological evaluations were carried out in laboratory rats exposed to the different emissions scenarios. This paper details the statistical approaches used to analyse the TERESA health data. Primary analyses focus on the effect of each exposure scenario at each power plant on each health outcome, by comparing the response means between exposed and unexposed animals, and the differences among these effects.

An additional level of analyses focuses on the associations between pollutant concentrations and health. The generated exposure scenarios at each of three separate power plants make statistical analyses of these associations, which relate to the health effects of a complex chemical mixture, challenging. Several statistical methods exist for assessing the effects of complex mixtures. These analytic strategies include (1) a multiple regression model that includes constituents as predictors, often called tracer elements, that are representative of meaningful constituent groups, (2) principal components regression, and (3) a two-stage analysis approach that first regresses an outcome on each constituent individually and then in a second stage regresses the resulting constituent slopes against indicator variables representing meaningful constituent groups. All of these approaches rely on some form of grouping of the constituents. For instance, the first and third approaches require the user to define constituent groups *a priori*, whereas principal components regression forms groups of constituents based on the correlation (or covariance) structure of the constituent concentrations. In concentrated ambient particle (CAPs) studies, for instance, these approaches have found wide application as the constituent groupings are taken to represent main particle sources.

In contrast, the TERESA exposure scenarios are generated from a single pollution source, coal-fired power plants, but differ due to the different mixes of primary and secondary particles formed through atmospheric reactions, as well as through the addition of atmospheric constituents (ammonia and secondary organic aerosol). Table 1 presents the exposure scenarios generated by this strategy. The analytic approaches listed above are not directly applicable for the TERESA exposure scenarios, because constituent groupings do not have a straightforward interpretation in terms of sources or another natural grouping. As a result, we propose a new approach to analyzing the health effects of a complex mixture. We first conduct univariate constituent association analyses, and then support these classical analyses with a novel multivariate analysis of exposure composition based on an extension of regression trees known as random forests. We illustrate our approach by analyzing data on the respiratory outcome PEF. This analysis demonstrates how the multiple levels of the analysis hierarchy complement each other to collectively identify constituents most strongly associated with health effects.

In addition to the complexity of the TERESA exposures, statistical analysis of the TERESA data must also account for the fact that the endpoints are recorded at different time resolutions. For instance, bronchoaleovolar lavage (BAL), blood, and chemiluminesence parameters are recorded once for each animal exposure, repeated electrocardiographic outcomes are analyzed on an hourly basis, and pulmonary function readings are recorded every 10 minutes during each animal's exposure. Accordingly, although we refer to these levels of exposure analysis as "ANOVA" and "regression", this strategy of increasing the sensitivity of the exposure metric is nested within regression extensions that respect the design of the study and the correlation structure of the data.

Section 2 outlines the exploratory data analysis strategies undertaken for each outcome. Section 3 outlines models for outcomes recorded once per animal, and Section 4 outlines the models used to analyze the electrocardiographic data. Section 5 outlines the analogous models for the respiratory outcomes measured semi-continuously during the exposure period. Section 6 describes the approach taken to analyze associations between health and

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

individual constituents. Section 7 reviews random forests, and outlines the application of this multivariate approach to the TERESA data. Section 8 describes methods for model comparison, and Section 9 illustrates the entire hierarchical approach to data analysis by applying the multiple levels of analysis. Throughout these sections, we refer to animals exposed to the generated aerosol as the "exposed animals", and animals exposed to filtered air as the "control animals". One important aspect of the TERESA study design is that control animals were assessed for each exposure occasion. Therefore, because responses from exposed animals are always compared to those from the corresponding controls, the fact that the number of animals assessed varied both by exposure scenario and plant (Godleski 2010, this series) affects the precision of the effect estimates but does not produce bias in these estimates.

2. EXPLORATORY ANALYSES

Initial analysis of each outcome included univariate explorations of all variables, using histograms, statistical summaries, and other graphical techniques. Expected ranges for all of the variables were defined a priori, and out of range values, or values identified as outliers were checked for errors. In addition to data cleaning, these initial graphical checks served the purpose of identifying skewed variables that might require transformation.

We took the conservative approach that primary analyses were based on all data. In limited cases, we ran secondary analyses after eliminating clear outliers in exposure concentrations due to the fact that such observations have extremely high leverage and may dominate regression results. This approach confirms the primary analyses by eliminating the possibility that any associations seen are driven by a few extreme observations.

3. INTEGRATED DATA: Bronchoaleovolar Lavage and Chemiluminescence

Bronchoaleovolar lavage (BAL), blood, and chemiluminescence (CL) data were recorded on a given animal at single timepoint after exposure. We assessed how these differences due to exposure varied by exposure scenario. These analyses treat exposure as a binary indicator of whether an animal was exposed or not, and scenario as a categorical variable.

Exploratory analyses showed that the distribution of the chemiluminescence outcomes were log-normally distributed. Therefore, before analyses, each reponse was log-transformed to satisfy the normality assumptions in regression analyses outlined below. As a result, regression parameters for this outcome represented log multiplicative effects of exposure, and we reported the percent change in this outcome associated with exposure. In addition, exploratory analyses also revealed significant day-to-day variability among control animals. Therefore, in models for these outcomes, we blocked on day. This corresponds to comparing response levels in exposed animals to those in the filtered air controls *from the same day*.

More specifically, let Y_i denote a biologic response for animal i, let e_i denote a binary indicator of whether that animals was exposed, let d_i be a vector of indicator variables for each day in the study, and let $I(x)_i$ be an indicator function that is equal to 1 if animal i was exposed under scenario x and 0 otherwise. For each BAL, blood, or CL outcome, we first fit the global exposure model

$$Y_i=\beta_0+\beta_1 e_i+\gamma d_i+\varepsilon_i, \quad (1)$$

with $\varepsilon_i:N(0,\sigma^2)$. Here, γ is a vector of regression coefficients capturing the day-to-day variability in response among control animals and d_i is a vector of indicator variables associated with the categorical *day* variable.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

If the estimate of β_1 in model (1) is signficantly different from zero, we refine the model to specify scenario-specific effects of exposure. Let S_i denote a 7×1 vector of indicator variables representing whether a given day corresponds to *P*, *PO*, *POS*, *PONS*, *O*, *OS*, *S* exposure. This model is

$$Y_i=\beta_0+\beta S_i e_i+\gamma d_i+\varepsilon_i, \quad (2)$$

In this characterization, β is a vector of scenario-specific exposure effects. An ANOVA F test comparing the fits of models (2) and (1) yields a test of whether the effect of exposure differs among the exposure scenarios. Exposure effects were assessed both on a plant-specific basis and overall.

Our interpretation of the results focuses on both the estimated magnitude of the estimated health effects, as well as the estimated strength of these associations. For the ANOVA models that assess the effects of exposure for each scenario, based on a Bonferroni correction for multiple comparisions we consider p-values less than 0.007 as strong evidence of an exposure effect for a given scenario, and a p-value satisfying $0.007 < p < 0.05$ as marginal evidence of an effect. That is, we consider the strength of evidence of an effect. We note that we applied this correction to the ANOVA analyses, but not the composition analyses.

4. ELECTROCARDIOGRAPHIC DATA

Cardiac arrhythmias and autonomic function were assessed under the POS exposure scenario at Plants 2 and 3. For a complete description of the plants, their coals, and their constituent levels, by scenario, see Kang et al. (2010 this series). We tested the hypothesis that exposure to POS increases premature ventricular beat (PVB) frequency (expressed as number of PVBs per hour, for each of five hours of exposure) as compared to control animals exposed to filtered air. We used repeated-measures Poisson regression (Diggle et al. 2002) fit using generalized estimating equations (GEE) to model PVB frequency during each exposure hour. This model included indicator variables for time (exposure hour), group (exposed versus filtered air), and 2-way interactions between these variables. We allowed for Poisson overdispersion in the data, assumed an exchangeable covariance structure, and based inferences on empirical (robust) standard errors. As a sensitivity analysis, we tested the related hypothesis that exposure to POS increases the risk of observing one or more PVBs as compared to control animals exposed to filtered air. We used repeated-measures logistic regression to model the odds of a given rat having one or more PVBs during a given time period. As above, this model was fit using GEE and included indicator variables for time, group, and the 2-way interactions between these variables. An exchangeable covariance structure was assumed and inferences were based on empirical (robust) standard errors.

We applied linear mixed models to one-hour averages of heart rate, heart rate variability parameters and interval data. The models included indicator variables for time, group, and the 2-way interactions between these variables as fixed effects and random rat-specific intercepts. Because these assessments were made in the POS scenario, consisting of four exposure occasions, it was not possible to assess associations between POS mass or constituent concentrations with any of the electrocardiographic outcomes.

5. SEMI-CONTINUOUS DATA: RESPIRATORY OUTCOMES

For 10-minute averaged respiratory function data, in which the data represent repeated measures on each animal throughout the six-hour exposure period, additive mixed models were applied to estimate both overall and time-varying effects of exposure. A form of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

repeated measures model for longitudinal data, additive mixed models (Coull, Schwartz, and Wand 2001; Ruppert, Wand, and Carroll 2003) represent an extension of linear regression models that allows one to (1) estimate potentially non-linear effects of independent variables, and (2) include random effects as independent variables in order to account for correlation among repeated measurements taken on the same animal during the exposure period. For each outcome, additive mixed models were fit using as independent variables (1) a general nonlinear mean trend for control animals over the exposure period, (2) an exposure indicator, which implies a constant shift in the mean trend due to pollutant exposure, and (3) random animal effects reflecting animal-to-animal heterogeneity that results in correlation among 10-minute averages taken on the same animal over time.

Exploratory analyses of the respiratory outcomes indicated some outcomes were normally distributed and some were log-normally distributed. However, because the large amount of data for these outcomes and the central limit theorem protects against misspecification of the normal assumption in regression analyses for these outcomes, it is likely that inferences from regression models fit to the untransformed outcomes are valid. To check this assertion, we re-ran the analyses on the respiratory outcomes after log-transforming the pulmonary outcomes, and the overall conclusions of the analysis did not change. The refore, for consistency of interpretation across respiratory outcomes, all models were applied to the data on the original scale. Sensitivity analyses were then run on the log-transformed values to confirm that these inferences were qualitatively similar, and this was indeed the case.

In order to define the specific models, let Y_{it} denote the response for animal i at time t. We assessed whether or not there was evidence that differences between exposure groups varied within the period of exposure. This model specified distinct mean trends over the exposure period for the exposed and filtered air animals, again including random animal effects to account for the repeated measurements taken on each animal. The difference between these estimated trends represents the time-varying effect of the exposure scenario over the exposure period. We applied this approach to each week's data separately using the model

$$Y_{it}=\beta_0+s_{e_i}(t)+b_i+\varepsilon_{it}, \quad (3)$$

where here $s_1(t)$ denotes the average trend among exposed and $s_0(t)$ represents the average trend among unexposed, and $\delta(t) = s_1(t)-s_0(t)$ represents the time-varying effect of exposure for a given week. The random subject effects $\{b_i\}$ account for unexplained subject heterogeneity that induces correlation among repeated responses taken on the same subject, above and beyond that explained by the smooth temporal trends $s_1(t)$ and $s_0(t)$.

Seeing little evidence of time-varying effects of exposure within an exposure period, we simplified model (3) to specify a constant effect of exposure over time. Like the analogous models for BAL and CL, we estimate within-day effects of exposure by controlling for day in the model. The model is

$$Y_{it}=\beta_0+\beta_1 e_i+\gamma d_i+s(t)+b_i+\varepsilon_{it}, \quad (4)$$

where again γ represents the nuisance day-to-day variability in the outcome among the control animals, b_i again represents animal-specific random effects accounting for animal-to-animal heterogeneity in the response that persists across the entire exposure period, and $s(t)$ is a smooth function of time that accounts for autocorrelation among measures taken on the same animal within an exposure. Like the integrated analyses, we expanded the models by replacing the overall exposure term with $\beta S_i e_i$. The models were fitted to data from each plant separately as well as to data from all three plants together.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

6. UNIVARIATE CONSTITUENT ASSOCIATIONS

To assess univariate associations between particle or gas concentrations and toxicological endpoints, we conducted single-pollutant analyses in which separate regression models were fitted using biologic response as the dependent variable and either mass, particle number, or a single concentration as the exposure metric. As noted in the Introduction, existing two-stage approaches do not directly apply to the exposure scenarios defined by the TERESA study, as plausible groupings of constituents are not necessarily constant across these different scenarios.

For outcomes for which there exist significant mean differences between exposure groups, an issue that arises is that most concentrations that are entered into a model univariately will show significant differences, due to the fact that there is high collinearity between concentration and exposure group. That is, it can be the case that differences between groups, and not a dose-response between concentration and outcome within the exposed group, drive an association between outcome and concentration. Panel (a) of Figure 1 shows an hypothetical example of such an instance. Suppose the four data points in this plot represent response means from a single week of exposure, with the point at zero concentration representing the average filtered air response for that week and the three points at non-zero concentrations representing the average responses among exposed animals on the three days. In this case, there is a strong effect of exposure, but no effect of increasing concentration among exposed animals, and the least squares fit of the regression model that contains concentration alone suggests an association between response and exposure.

Accordingly, we used as a more stringent definition of an association between a particular exposure constituent and outcome. We require this association to be present among animals in the exposed group. If we control for day-to-day differences in response in control animals, this leads to the model

$$Y_i=\beta_0+\beta_1 e_i+\beta_2 c_i+\gamma d_i+\varepsilon_i, \quad (5)$$

where c_i is the concentration of the particular constituent of interest to which animal i was exposed. Here we assume all control animals receive $c_i = 0.0$. This is not unlike well-established methods to estimate the effects of smoking (among never, former, and current smokers) in epidemiological analyses. Such methods typically include both an indicator term for (former and current smokers), reflecting that there is some baseline effect of any exposure, as well as the number of cigarettes smoked by a current smoker, which reflects the effect of the amount of current exposure.

For the purposes of convenient visualization of these univariate associations, we sum both sides of the underlying model (4) to yield a simple model for the daily differences in group means as a function of daily constituent concentrations. Let $\bar{Y}_d^E$ and $\bar{Y}_d^C$ denote the mean outcome for the exposed and control groups on day d, respectively, and define $Z_d=\bar{Y}_d^E-\bar{Y}_d^C$ to be the difference between these group means. On the daily scale, model (4) equates to

$$Z_d=\alpha_0+\alpha_1 c_d+\eta_d, \quad (6)$$

where c_d is the constituent concentration of the generated aerosol for day d. That is, this model reduces to a simple linear regression for these daily differences. This suggests that we can plot these daily differences against the daily concentrations as an easy way to graphically depict these associations. We report the estimated univariate associations and

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

corresponding standard errors by standardizing α_1 by the standard deviation of that concentration, so that the resulting estimates represent the change in daily group difference for a given standard deviation increase in that constituent concentration.

In analyzing these univariate associations, we focused both on the direction and magnitude of the associations for each constituent. We interpreted associations in which low levels of a constituent were associated with small changes in outcome and high levels were associated with large changes, rather than vice versa, as consistent with a causal hypothesis for that constituent. Conversely, cases in which low concentrations of a constituent are associated with large differences in outcome means and vice versa suggest that, assuming the result is not a false positive (Type I error), the particular constituent may simply be inversely correlated with a causal agent. Moreover, we checked whether or not such observed associations are consistent with the scenario-specific ANOVA analyses, further ensuring that the observed findings are internally consistent. Finally, we inspected scatterplots of each outcome difference, color coded by plant and symbol coded according to scenario, to investigate how the ANOVA and univariate constituent models complement each other.

7. MULTIVARIATE ANALYSES

The standard approach to assessing the multivariate association between a health outcome and multiple exposure constituents is a multivariate regression with main effects for each constituent. However, there are a few disadvantages to this approach. First, it typically relies on p-values from a hypothesis testing framework to determine strength of association, which can be subject to multiple testing issues when applied to a large number of constituents. Second, the true model, in terms of which variables should be included in the model, is uncertain. Third, such a model with constituent main effects only makes an assumption of additivity for the effects of the included constituents, in that no constituent-by-constituent interactions are entered into the model. Fourth, it assumes that the concentration of each constituent concentration is linearly related with the outcome.

Accordingly, we used a relatively new statistical method known as random forests (Breiman 2001; Liaw and Wiener 2002) to rank constituent concentrations on their ability to predict mean differences in a biologic outcome on a given day. Random forests are an extension of the relatively well-established approach of classification and regression trees (CART; Hastie et al. 2003), which is an inherently nonlinear regression framework that partitions the complex predictor space into regions thought to have common values of the outcome. Brieman (1996) showed that repeatedly applying the CART methodology to random subsets of the data can have several advantages over the standard CART approach that builds only a single tree. First, conclusions are less sensitive to the size of a single tree, as results are averaged over multiple trees. Second, this extension can achieve lower prediction error than single tree results. In the following two subsections, we outline the basic idea of a regression tree and its recent extension to random forests and variable importance.

7.1 Background: Regression Trees

Regression trees represent a method for clustering a multivariate predictor, guided by associations between the predictors and an outcome. That is, regression trees represent a form of hierarchical clustering in which the sample is successively split into groups defined by dichotomizing a predictor into low and high values. The predictor that is chosen for splitting as well as the value of the predictor at which the split is made is selected so as to maximize the difference in the mean values of the outcomes in the resulting groups. Figure 4 presents a diagram of this process when interest focuses on a two predictors, x_1 and x_2. The regression first splits the sample according to a particular value, a of x_1, chosen to maximize the difference in outcome means for these two subgroups. The algorithm then goes on to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

split two resulting subsamples further, splitting the first subsample according to whether or not the value of x_2 is less than b and the second subsample according to whether or not the value of x_2 is less than c. The resulting clusters are assumed to consist of observations with a common mean. This mean is estimated using the sample mean of the outcome for observations contained in that cluster.

For the TERESA study data, we applied regression trees at the daily level; that is, we used as the endpoint Z_d, the difference between outcome means in the exposed and unexposed groups on day d. Figure 5 shows the hierarchical clustering sequence when we apply the algorithm to the PEF mean differences. The figure shows that the data were first split according to whether or not the elemental carbon (EC) concentration on that day was less than 0.2 $\mu g/m^3$. Ammonium ion and acidic sulfate (SO_4) were the constituents on which splitting was based at the second level, and so on.

Note that the regression tree framework allows for a flexible association structure between an outcome and a multivariate predictor. First, the model does not assume a linear exposure-response relationship between outcome and a given constituent, due to the categorization of each predictor. Second, unlike standard multiple regression analyses containing a main effect for selected constituents, regression trees allow for synergy or other complicated interactions among constituents. This is possible because each constituent is a candidate for the basis of splitting for each subgroup generated by a previous split. So, for the PEF analysis presented in Figure 5, acidic SO_4 appears to be a more important predictor for those samples for which EC concentration is high, but not for samples for which EC concentration is low.

7.2 Random Forests

The advantages of regression tree analyses are balanced by a few disadvantages. First, one must decide how many levels to run the analysis. One approach is to sequentially split the sample many times, and then “prune” the tree and remove levels according to one of several optimality criteria. Second, regression trees do not provide a direct measure of the overall importance of a given predictor in predicting the outcome. A recently developed extension of regression trees, random forests, addresses these issues.

The random forest algorithm repeatedly applies the regression tree algorithm to bootstrap resamples of the original data set. The basic idea of the algorithm is to use the data not included in each bootstrap to serve as a “out-of-sample” dataset on which prediction accuracy of a given tree can be assessed. Regression trees assess the importance of a given variable in predicting the outcome by permuting the predictors, and assessing how much prediction accuracy decreases after this permutation. Formally, the random forests algorithm is as follows:

1. Draw n_{tree} bootstrap samples from the original data.
2. For each of the bootstrap samples, grow an unpruned regression tree, but with a random constituent: at each node of the tree, rather than choosing the best split in the response among all constituent predictors, randomly choose m of the predictors and select the best split among these m constituents.
3. An estimate of the prediction mean squared error can be obtained by predicting the data not in each bootstrap sample (denoted “out-of-bag”, or OOB, data) using the tree grown with the bootstrap sample. These predictions for each data point are then averaged over all of the trees for which that point is out-of-bag. On average, each data point is in the OOB data approximately 36% of all samples.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4. Compute the "Variable Importance" of each predictor. The importance of a predictor in the random forest algorithm is estimated by the increase in the prediction error when the value of that predictor is randomly permuted in the OOB data, leaving all other variables unchanged. More specifically, let d_t be the difference in OOB prediction mean squared error (IncMSE) with and without permuting the variable of interest, for the tth tree. The overall variable importance measure is mean of $d_t/se(d_t)$, where $se(d_t)=sd(d_t)/T$ (%IncMSE).

As noted above, we apply random forest methodology to the TERESA data at a daily level. That is, the response is Z_d, the difference in means from the exposed and non-exposed groups and the predictors are the measured constituent concentrations for that day. All random forest analyses were run using the *RandomForest* package in the R statistical software package (Liaw and Wiener 2002).

Each random forest analysis is determined by two user-specified parameters, the number of bootstrap samples (n_{tree}) and the number of predictors chosen in each resample (m). For each biologic outcome, we selected n_{tree} = 50,000. For m, we followed the recommended default (Breiman 2001) of $p/3 \approx 9$, where p is the total number of constituents in the analyses. For each outcome we also performed a sensitivity analyses that re-ran the analysis with n_{tree} = 100,000, and with m = 4 and m = 18. In each case, although the precise estimate of the MSE increase associated with a random permutation of a constituent's concentrations varied across these sensitivity analyses, the relative ordering of the importance of each constituent was quite stable across these multiple runs (see Liaw and Weiner 2002 for a discussion of this point).

8. MODEL COMPARISON

An important question that arises is whether model with plant + exposure scenario or the model incorporating constituent concentrations does a better job predicting effects of exposure. Accordingly, we use R^2 comparisons to compare the (model-size adjusted) performance of (1) the ANOVA models allowing the effects of scenario to vary by plant, (2) univariate models for a subset of important constituent, (3) a multivariate model incorporating all constituent data via a random forests analysis. All models are applied to Z_d, the daily mean differences in outcome.

We use two forms of R^2 to make these comparisons. First, we used adjusted R^2_{adj}, defined as $R^2_{adj}=1-\frac{N-1}{N-p-1}\frac{SSE}{SST}$ where SSE and SST are the error and total sums of squares, respectively, p is the number of parameters in the model and N is the total sample size. Because our chosen multivariate method of analysis, random forest analysis, assesses predictive accuracy, we also compared the models based on predictive R^2. For the random forest analyses, this measure is estimated using the "out-of-bag" sample for each bootstrap resample, and averaged over all re-samples. For the ANOVA and univariate regression models, we estimated predictive R^2 using 10-fold cross-validation. Ten-fold CV sets aside one-tenth of the sample, fits a model to the remaining data, calculates the squared error of these predictions for each observation in the held-out data, and repeats this process across each tenth of the data. As noted in the Introduction, we view the random forest approach as a flexible model allowing a wide range of exposure-reponse relationship between each constituent concentration and response, and the possibility of interactions among constituents. Therefore, the comparison of the predictive R^2 comparison compares the plant +scenario model to one of the least-restrictive constituent models, and therefore speaks directly to whether the constituent data add predictive power above that provided by the simpler plant/scenario classifications.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

9. ILLUSTRATION

To illustrate the approach taken to analyze the TERESA toxicological data, we continue with the example of peak expiratory flow. Because PEF is measured semi-continuously, the ANOVA models represent a linear additive mixed model. We fit model (3) at the finest level of resolution, estimating a time-varying effect of exposure within each week of experiments. Figure 2 shows these plots (averages over a 4-day period) for the exposures at Plant 2 during the period of May 9 - 12, 2006. The results do not provide much evidence of time-varying exposure effects, and analogous plots from other weeks (data not shown) similarly showed little variation in effect within an exposure period. Model (4) estimated the average effect of exposure over time. Stratified by power plant (Table 2), this effect was strongest at Plant 2, partially present at Plant 1 (only for PONS), but was not observed at Plant 3. The interaction p-value tests the null hypothesis that the exposure effect is constant across scenarios, versus the alternative that they are different across scenarios. Table 3 shows this effect broken down by scenario, pooled across all three power plants. The results suggest that, for this outcome, the PONS and P scenarios are associated with the largest effects on PEF.

For easy visualization of associations between the outcome and individual constituents of exposure, and how these univariate associations are driven by differences in outcome among plants and among scenarios, univariate association analyses are based on model (5). Figure 3 shows the association between the PEF mean differences and elemental carbon concentration, color coded by plant and symbol coded by exposure scenario. The p-value corresponding to the hypothesis that the slope of this association is zero is $p = .017$. The color coding suggests that plant is an important factor driving this association between PEF difference and EC, as the PEF values from Plant 3, with EC concentration of 0, are higher than those from Plants 1 and 2.

Figure 6 shows a plot of the estimated variable importance for the constituent predictors from a multivariate random forest analysis of the PEF data. This figure shows that EC concentration has highest variable importance score among all constituents, suggesting that the association shown in Figure 3 is the strongest among all constituents, and cannot be explained by the presence of the other constituents entered into the multivariate model.

We compared the performance of the ANOVA and univariate regression models using adjusted R^2. Table 4 presents the adjusted R^2 for the the two-way ANOVA model using plant, scenario, and their interaction as well as for several selected single constituent regression models. The results suggest that the model using plant and scenario information explain variation in exposure effects by an order of magnitude more than that based on any single constituent. We use predictive R^2 to compare the predictive performance of the plant/scenario model to the random forest model, which incorporates all constituent data in a multivariate fashion. Note that because these are out-of-sample R^2 values, it is possible that they can be negative.These values, also presented in Table 4, suggest that these models have similar predictive power, suggesting that incorporation of the constituents listed in Figure 6 does not improve upon the plant/scenario model.

10. DISCUSSION

Statistical analyses of the data arising from the TERESA study presented unique challenges related to assessment of the health effects of complex mixtures. In this paper we consider ANOVA, univariate regression, and random forests as a broad analysis of the relative importance of plant, scenario, and individual constituents in explaining exposure effects.

In interpreting the results from the different levels of analysis for a given outcome, we looked for consistency across analysis levels as well as associations consistent with a causal

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

hypothesis. For instance, in the illustrative case study of PEF we presented here as well as in Diaz et al. (2010 this series), the ANOVA-type analyses suggest the PONS and P scenarios yield the largest effects on respiratory function. Stratification by plant suggested that these effects varied largely by plant, with Plant 2 exhibiting relatively large effects for these scenarios, Plant 1 showing effects only for the PONS scenario, and Plant 3 not showing any effect on PEF. Univariate regression analyses suggest that these differences are most strongly associated with EC concentrations. This observation is consistent with the plant-specific analyses, as EC concentrations vary across plants. An R^2 analysis confirms that scenario and plant drive these effects much more so than the constituent concentrations for this outcome.

Random forests have seen increasing use in high-dimensional genomic settings (e.g., Ye et al. 2005). To our knowledge there have been relatively few applications of this method of analysis in the environmental sciences, and environmental toxicology in particular. A natural line of future research is a comparison of the operating characteristics of this approach when interest focuses on identifying important constituents of a complex exposure mixture, and comparing this performance to existing approaches to this problem.

Acknowledgments

This project was supported by the Electric Power Research Institute (Contract EP-P10983/C5530/56546), the U.S. Environmental Protection Agency Center, for Particle Health Effects at the Harvard School of Public Health (grant R827353), and the Harvard NIEHS Center for Environmental Health (grant ES00002). This work was also prepared with the support of the U.S. Department of Energy (DOE) under award DE-FC26-03NT41902, and a grant from the State of Wisconsin. However, any opinions, findings, conclusions, or recommendations expressed herein are those of the authors, and do not necessarily reflect the views of the U.S. EPA or the DOE. The authors thank members of the TERESA External Advisory Committee for helpful comments that improved the quality of the manuscript.

REFERENCES

Breiman L. Random forests. Machine Learning. 2001; 45:5–32.

Capanu M, Orlow I, Berwick M, Hummer AJ, Thomas DC, Begg CB. The use of hierarchical models for estimating relative risks of individual genetic variants: An application to a study of melanoma. Statistics in Medicine. 2008; 27:1973–1992. [PubMed: 18335566]

Coull BA, Schwartz J, Wand MP. Respiratory health and air pollution: Additive mixed model analyses. Biostatistics. 2001; 2:337–349. [PubMed: 12933543]

Diaz EA, Lemos M, Long M, Coull BA, Ruiz PA, Gupta T, Kang CM, Vlassidis E, Godleski JJ. Toxicological Evaluation of Realistic Emission Source Aerosols (TERESA): Breathing Pattern Health Effects Related to Power Plants. Inhal Toxicol. 2010 Submitted as part of this series to.

Diggle, PJ.; Heagerty, PJ.; Liang, K-Y.; Zeger, SL. Analysis of Longitudinal Data. 2nd Ed. University Press; Oxford: 2002.

Godleski JJ, Rohr AC, Kang CM, Diaz E, Ruiz PA, Koutrakis P. Toxicological Evaluation of Realistic Emission Source Aerosols (TERESA): Introduction and Overview. Inhal Toxicol. 2010 Submitted as part of this series to.

Hastie, T.; Tibshirani, R.; Friedman, JH. Elements of Statistical Learning. Springer; New York: 2001.

Kang C-M, Gupta T, Ruiz PA, Wolfson JM, Ferguson ST, Lawrence J, Rohr AC, Godleski JJ, Koutrakis P. Aged particles derived from emissions of coal-fired power plants: the TERESA field results. Inhal Toxicol. 2010

Liaw A, Wiener M. Classification and Regression by random Forest. R News. 2002; 2(3):18–22. Available at http://CRAN.R-project.org/doc/Rnews/.

Ruppert, D.; Wand, MP.; Carroll, RJ. Semiparametric Regression. Cambridge University Press; Cambridge, UK: 2003.

Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. Informative Bayesian structural equations models for assessing source-specific health effects of air particulates. Biostatistics. 2007; 8:609–624. [PubMed: 17032699]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Nikolov MC, Coull BA, Catalano PJ, Diaz E, Godleski JJ. Statistical methods to evaluate health effects associated with major sources of air pollution: A case study of breathing patterns during exposure to concentrated Boston air particles. Jl Roy Statist Soc, Ser C. 2008; 57:357–378.

Lewinger JP, Conti DV, Baurley JW, Triche JT, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. Genet Epidemiol. 2007; 31:871–882. [PubMed: 17654612]

Thomas DC, Witte JS, Greenland S. Dissecting effects of complex mixtures - Who's afraid of informative priors? Epidemiology. 2007; 18:186–190. [PubMed: 17301703]

Ye Y, Zhong X, Zhang H. A genome-wide tree- and forest-based association analysis of comorbidity of alcoholism and smoking. BMC Genetics. 2005; 6:S135. [PubMed: 16451594]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

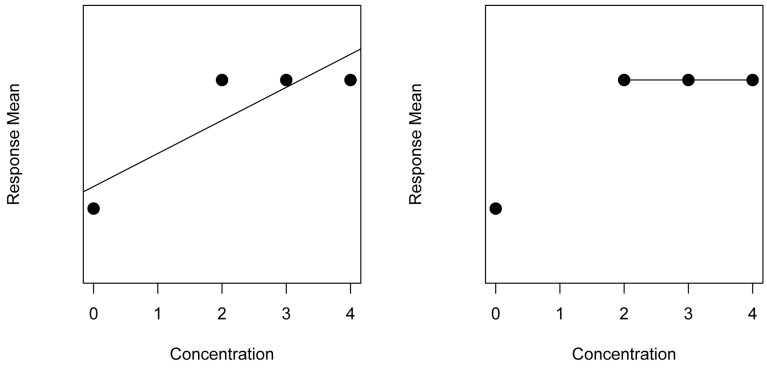

Figure 1.
Hypothetical Constituent-Response Association

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

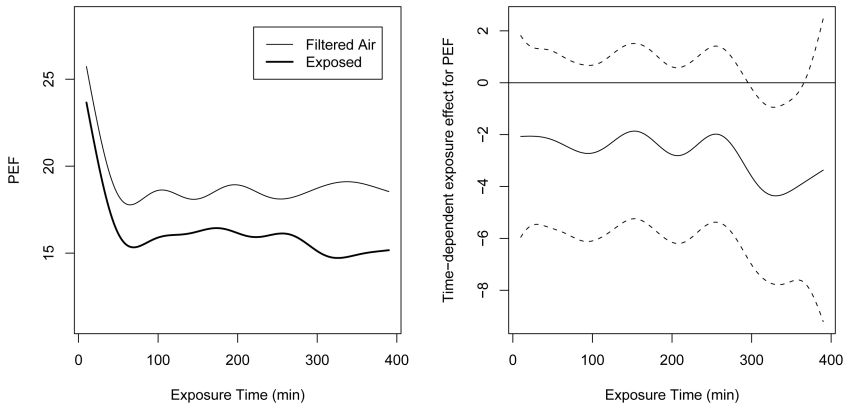

Figure 2.
Estimated Time-varying Exposure Effect on PEF (in ml/s) During May 9-12, 2006

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

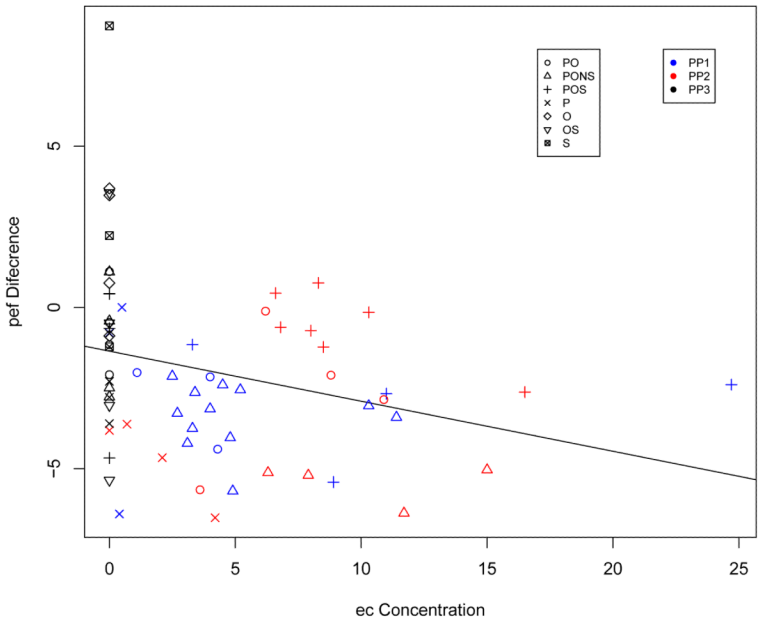

Figure 3.

Univariate Association Between Daily PEF Mean Difference (from controls) and Elemental Carbon Concentration, by scenario and plant.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

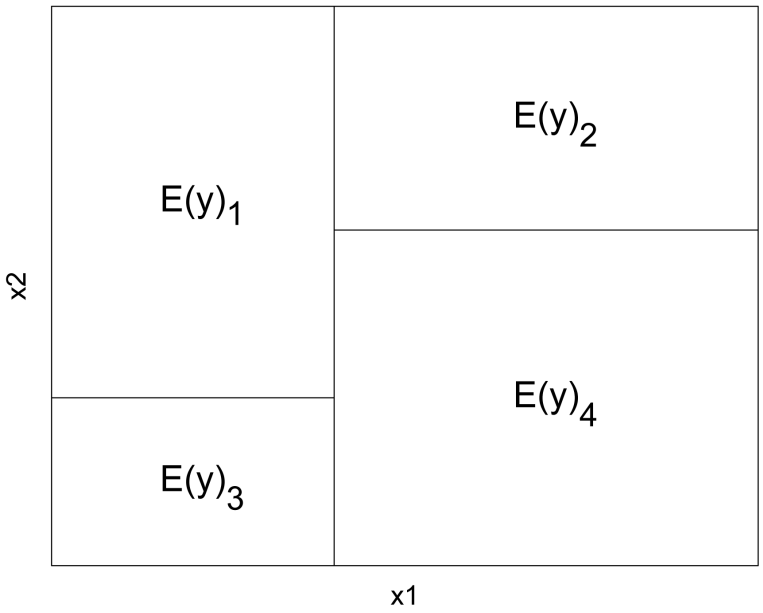

Figure 4.
Diagram of regression tree partitioning of 2-dimensional predictor space

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

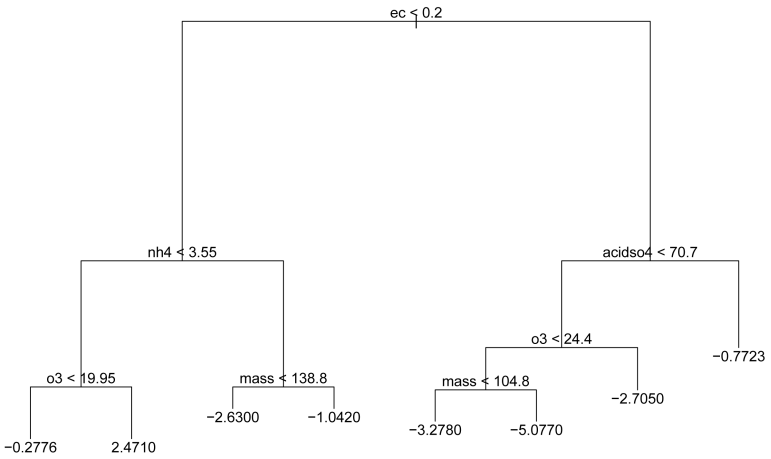

Figure 5.
Results of regression tree modeling for daily mean differences of peak expiratory flow outcome

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

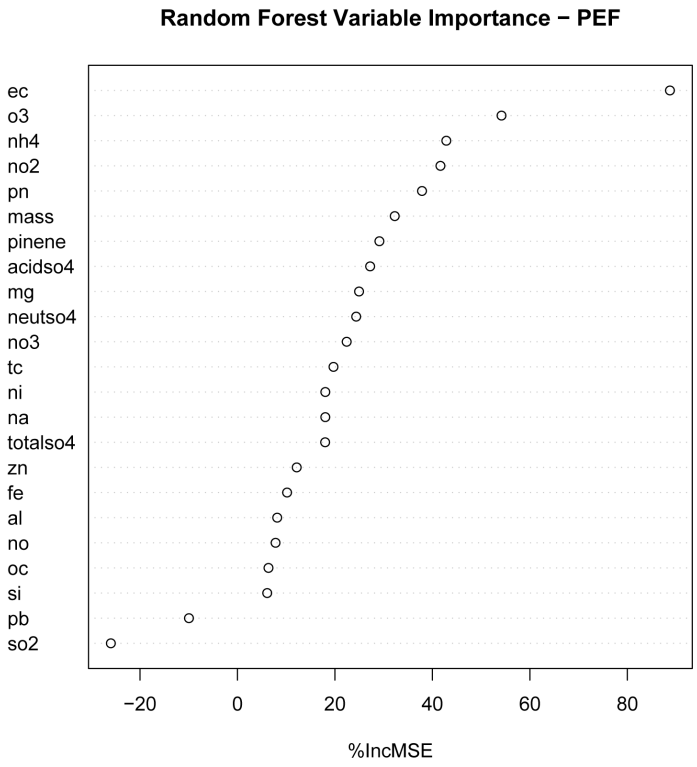

Figure 6.
Random forest outcome for peak expiratory flow

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

TERESA Exposure Scenarios

Scenario	Description
P	No-aging primary emissions only
PO	P Scenario + H_2SO_4 aerosol from oxidation of SO_2
POS	PO Scenario + reaction of α-pinene with ozone to form SOA
PONS	POS Scenario + neutralization of H_2SO_4 aerosol by NH_3
O	Oxidation of SO_2 including primary gases, but not primary particles
OS	O Scenario + SOA
S	SOA produced using no primary gases or particles

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Estimated Plant- and Scenario-specific Exposure Effect (in ml/s) on PEF

	PP1			PP2			PP3		
	Exp - FA			Exp - FA			Exp - FA		
Scenario	**Diff**	**SE**	**p-value**	**Diff**	**SE**	**p-value**	**Diff**	**SE**	**p-value**
PONS	−3.36	1.12	0.003	−5.45	1.20	< 0.0001	−1.13	2.75	0.681
POS	−2.91	1.95	0.135	−0.61	1.01	0.551	−1.12	2.75	0.683
PO	−2.85	2.25	0.205	−2.68	1.20	0.026	−0.65	2.57	0.800
P	−2.27	2.25	0.312	−4.66	1.20	0.0001	−2.26	2.87	0.431
OS							−1.33	2.74	0.628
O							1.76	2.57	0.494
S							3.26	3.16	0.302
Interaction p-value	0.976			0.010			0.842		

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

Estimated Scenario-specific Exposure Effect on PEF (in ml/s), Across All Plants

	Exp - FA		
Scenario	**Diff**	**SE**	**p-value**
PONS	−3.5270	0.8464	< 0.0001
POS	−1.4753	1.0492	0.1597
PO	−2.1620	1.1641	0.0633
P	−3.3054	1.1984	0.0058
OS	−1.3316	2.3422	0.5697
O	1.7587	2.1927	0.4225
S	3.2618	2.7013	0.2273
Interaction p-value	0.081		

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 4

Adjusted and Predictive R^2 for ANOVA, univariate regression, and random forest models for PEF mean differences

Model	Adjusted R^2	Predictive R^2
Plant/Scenario ANOVA	0.43	0.10
Neutralized SO_4	0.09	0.06
Acidic SO_4	0.03	0.00
Total SO_4	0.00	−0.08
Mass	0.01	−0.01
EC	0.07	0.04
NH_4	0.11	0.11
Random Forest	–	0.09