

## APPENDIX F

VASCONCELOS, L. A. DE P., MACIAS, E., McMURRAY, P. H., TURPIN, B. J., AND W. WHITE, 1999. A CLOSURE STUDY OF EXTINCTION APPORTIONMENT BY MULTIPLE REGRESSION, SUBMITTED TO *ATMOSPHERIC ENVIRONMENT*, JULY

# A CLOSURE STUDY OF EXTINCTION APPORTIONMENT BY MULTIPLE REGRESSION

L.A. de P. Vasconcelos<sup>1</sup>, E.S. Macias<sup>1</sup>, P.H. McMurry<sup>2</sup>, B.J. Turpin<sup>3</sup>, and W.H. White<sup>1\*</sup>

1. Chemistry Department, Washington University, St. Louis, MO 63130
  2. Mechanical Engineering Department, University of Minnesota, Minneapolis, MN 55455
  3. Environmental Science Department, Rutgers University, New Brunswick, NJ 08901
- \* Author to whom correspondence should be addressed

7/99

## ABSTRACT

Multiple regression has been widely used to apportion particle light scattering among distinct chemical species. The resulting scattering budgets are shown here to be unbiased estimates under certain theoretical conditions. The theory allows species' particle size distributions and water uptakes to vary from sample to sample, as they are known to do in reality. The sole constraint is that variations in each species' characteristics be statistically independent of all species' concentrations. Individual violations of this condition cause identifiable biases, and multiple violations can offset each other to yield regression estimates that are accurate by accident. Detailed and summary accountings of statistical errors are illustrated for examples based on actual measurements.

## INTRODUCTION

We often wish to resolve a pollution effect into a sum of contributions from different types of emissions. The light scattering budget used to apportion visibility effects is a familiar example of such accounting (e.g. Sisler and Malm, 1994). A scattering budget represents scattering by particles,  $\sigma_{sp}$  ( $Mm^{-1}$ ), as a linear function of particulate species concentrations,  $x_j$  ( $\mu g/m^3$ ):

$$\sigma_{sp} = \sum e_j x_j. \quad [1]$$

The term  $e_j x_j$  represents the scattering attributed to the  $j^{\text{th}}$  species. The mass-specific scattering coefficient  $e_j$  ( $Mm^{-1}/[\mu g/m^3] = m^2/g$ ), often referred to as an "efficiency", depends on the species' refractive index, water uptake, and particle size distribution (White, 1986).

Of all air pollution effects, reduced visibility is the best understood at the level of

physical law. With sufficiently comprehensive measurements of individual particles, the scattering coefficient could be calculated directly from electromagnetic theory. Actual field measurements are of course incomplete chemically and/or aggregated over many particles. However, the most detailed of these measurements can be augmented by plausible and consistent physical assumptions to define model aerosols that do support theoretical calculations (Sloane, 1983; Zhang et al., 1994; Lowenthal et al., 1995). Such hybrid models express our present understanding of the functional relationship between visibility and atmospheric composition.

Multiple regression is sometimes substituted for theoretical modeling in visibility analyses (e.g. Appel et al., 1985). Most particle composition data are from filter measurements that aggregate particles of all diameters up to 2.5  $\mu\text{m}$  or more. These unresolved measurements are inadequate for theoretical calculations, because light scattering is a strong function of particle size. If actual scattering is also measured, however, its regression on bulk composition yields an empirical relationship similar to [1]:

$$\sigma_{\text{sp}} = \sum e_j^{\text{OLS}} x_j \pm \epsilon. \quad [2]$$

The scatter  $\epsilon$  in [2] is typically small, accounting for less than 10% of the variance in  $\sigma_{\text{sp}}$ . The regression coefficients  $e_j^{\text{OLS}}$  are interpreted as estimates for the specific scattering coefficients  $e_j$  (White, 1976; Anderson et al., 1994).

Although similar to each other in form, equations [1] and [2] represent different relationships. Equation [1] is a theoretical relationship, and holds exactly for individual aerosol samples in which distinct species are mixed externally, as distinct collections of particles. The specific scattering coefficients  $e_j$  for external mixtures can be calculated directly from particle characteristics, and describe a causal relationship between scattering and mass concentration. In contrast, equation [2] is an approximate description of multiple samples of aerosols in which species are arbitrarily mixed. The regression coefficient  $e_j^{\text{OLS}}$  describes an observed association between adventitious variations in scattering and concentration. White (1986) presented idealized examples of atmospheric fluctuations that would be unrepresentative of the perturbations expected from changed emissions.

This paper assesses the practical importance of the theoretical distinction between specific scattering and regression coefficients. The examples of White (1986) show that the atmosphere may respond to novel perturbations in ways not observed among existing fluctuations. One might still hope, however, that such hypothetical distinctions would somehow "average out" in the real atmosphere. The question may be framed as follows: how reliably do regression analyses of observed atmospheric data illuminate the underlying causal dependence of scattering on aerosol composition?

## METHODS AND DATA

Experimental comparisons between  $e_j$  and  $e_j^{\text{OLS}}$  can be difficult to interpret. *In situ* observations of  $e_j$  are available only for "species" that can be isolated for measurement, such as particle size ranges that one can sort aerodynamically or electrostatically (e.g. White et al., 1994). Calculation of  $e_j$  from particle measurements requires many assumptions (Lowenthal et al., 1995). The regression coefficients  $e_j^{\text{OLS}}$  are straightforward statistics of observables, but can be sensitive to measurement precision (White and Macias, 1987) and model selection (Sloane, 1988). Differences between  $e_j$  and  $e_j^{\text{OLS}}$  can arise from any of these factors in addition to the distinction of interest here, which is the distinction between atmospheric fluctuations and emissions perturbations.

McMurry and coworkers at the University of Minnesota Particle Technology Lab (PTL) have described aerosol models that integrate detailed particle measurements with electromagnetic theory (McMurry and Zhang, 1991; Zhang et al., 1993; Zhang et al., 1994; McMurry et al., 1996). Their work yields synthetic observation sets for which the exact relationship of scattering to particle composition is known *a priori*. Within the closed worlds of these models, both scattering and concentration "measurements" are made without error. More importantly, the assumptions on which specific scattering coefficients are calculated are known to be valid. The PTL models thus offer us a virtual laboratory in which  $e_j$  and  $e_j^{\text{OLS}}$  are directly comparable.

The next section identifies theoretical conditions under which classical regression analyses yield unbiased apportionments of light scattering to individual particle species. It goes on to derive an equation that relates errors in the regression estimates to violations of these conditions. The equation offers a "statistical microscope" that resolves the difference  $e_j^{\text{OLS}} - e_j$  into distinct components, each arising from atmospheric correlations that regression analysis neglects. A subsequent section illustrates this "microscopic" view with an example application to PTL model data.

In order to focus on statistical issues without distractions from particle-size dynamics (cf. White, 1986), attention in the next two sections will be restricted to external mixtures. The exact linearity of the scattering/mass relationship in this special case facilitates comparisons with the necessarily linear approximations produced by regression. The extension to aerosol models that include internal (within-particle) mixing is straightforward.

The examples considered in the present paper are based on comprehensive fine-particle

measurements from the Southern California Air Quality Study (SCAQS). SCAQS yielded complete data sets for 33 four-hour daytime samples and 11 twelve-hour overnight samples during the summer of 1987 at Claremont (McMurry and Zhang, 1991). Bimodal lognormal particle size distributions were determined from impactor data for organic and elemental carbon, sulfate, nitrate, and other ions, and iron. Aerosol water uptake was determined by tandem differential mobility analysis (TDMA) (McMurry and Stolzenburg, 1989).

In our examples, the aerosol is modeled as a mixture of nitrate, sulfate, carbonaceous, and soil particles (McMurry and Zhang, 1991). Each chemical fraction is given its observed particle size distribution and assigned unbound water inferred from TDMA. Scattering coefficients for the resulting particle distributions are calculated from electromagnetic theory for spheres (Bohren and Huffman, 1982). The calculations show reasonable agreement with measured scattering coefficients (not used here). Results are insensitive to details of the aerosol model, however, so the agreement with observations provides no support for external mixing in the actual aerosol (McMurry et al., 1996).

## THEORY

The energy scattered by a collection of particles is the sum of the energies scattered by individual particles. Aerosols whose constituents are partitioned into mutually exclusive classes of particles are thus natural subjects for scattering budgets. Such aerosols are described as external mixtures (Jaenicke, 1978).

Distinct species in an externally mixed aerosol consist of disjoint subsets of the aerosol's particles. Each subset can be regarded as a subaerosol, whose contribution to total scattering is the sum of the contributions from its member particles. The  $j^{\text{th}}$  species in an external mixture thus has a well defined scattering coefficient  $\sigma_j$ , and the total scattering is the sum of species contributions:  $\sigma = \sum \sigma_j$ .

The specific scattering of an externally mixed species  $j$  is defined as the ratio  $e_j = \sigma_j/x_j$  of species scattering to species mass. To distinguish specific scattering coefficients ( $e_j$ ,  $\text{m}^2/\text{g}$ ) more clearly from species' scattering coefficients ( $\sigma_j$ ,  $\text{Mm}^{-1}$ ), we shall often refer to the former as *efficiencies* (cf. White, 1986). Scattering is a strong function of particle size, and a species' distribution with respect to particle size responds to atmospheric processes and shifts in the mix of emissions. The efficiency  $e_j$  thus varies from sample to sample; we shall treat these variations as random fluctuations about a population mean, the expected value.

Efficiencies are modeled as random variables that are statistically independent of species concentrations. More precisely, let  $e_{ij}$  be the efficiency of species  $j$  in observation  $i$ , and let  $E(e_j)$  be the expected value of  $e_j$  in the sampled atmosphere. Let  $x_{ij}$  be the concentration of species  $j$  in sample  $i$ , and suppose that a series of  $n$  samples yields the matrix  $\mathbf{x} = x_{11}, \dots, x_{1k}; x_{21}, \dots, x_{2k}; \dots; x_{n1}, \dots, x_{nk}$  of observed concentrations. When we refer to the fluctuations in  $e_j$  as random, we mean that knowing all concentrations measured in all observations doesn't help to predict the efficiency  $e_{ij}$  associated with any individual observation. The conditioned expectation remains the unconditioned mean:

$$E(e_{ij} \mid \mathbf{x}) = E(e_j) \text{ for each } i \text{ and } j. \quad [3]$$

For notational economy the population mean  $E(e_j)$  will be indicated simply by  $e_j$  in the remainder of this paper, individual values  $e_{ij}$  from the population being distinguished by the added subscript.

Vasconcelos et al. (1994) reported an instance of conditions likely to fail condition [3]. Near the Grand Canyon in summer, they found a statistical association between high relative humidities and high sulfate concentrations. This suggests that  $E(RH_i \mid \text{high } x_{i,\text{sulfate}}) \gg 0$ . High humidities inflate sulfate scattering/mass ratios, because unbound water increases the scattering cross sections of hygroscopic particles without increasing measured species mass (Charlson et al., 1978). This means  $E(e_{i,\text{sulfate}} \mid RH_i \gg 0) > e_{\text{sulfate}}$ , suggesting  $E(e_{i,\text{sulfate}} \mid \text{high } x_{i,\text{sulfate}}) > e_{\text{sulfate}}$ , since  $E(RH_i \mid \text{high } x_{i,\text{sulfate}}) \gg 0$ . But  $E(e_{i,\text{sulfate}} \mid \text{high } x_{i,\text{sulfate}}) > e_{\text{sulfate}}$  is contrary to [3].

Following Cass (1979) and Trijonis (1979), many analysts have accounted for the effect of humidity on scattering by introducing a modified concentration,  $x_{ij,\text{wet}} = x_{ij}/(1-RH_i)$ . To the degree that  $e_{ij} = e_{ij,\text{dry}}/(1-RH_i)$  accurately models the functional dependence of scattering efficiency, the change of variable allows analysis to proceed in terms of a "dry" efficiency that is independent of humidity:

$$\sigma_{ij} = e_{ij}x_{ij} = [e_{ij,\text{dry}}/(1-RH_i)]x_{ij} = e_{ij,\text{dry}}[x_{ij}/(1-RH_i)] = e_{ij,\text{dry}}x_{ij,\text{wet}}.$$

The mean efficiencies  $e_j$  (wet or dry) provide an approximate description of scattering as a deterministic function of species concentrations:

$$\sigma_i = \sum_j e_j x_{ij} + \epsilon_i,$$

where

$$\epsilon_i = \sum_j (e_{ij} - e_j) x_{ij}.$$

By condition [3], the  $e_j$  suffice to apportion mean scattering. Letting  $\text{mean}_i$  denote the mean over all observations, we have:

$$\text{mean}_i[\sigma_{ij}] = \text{mean}_i[e_{ij}x_{ij}] \approx \text{mean}_i[E(e_{ij}x_{ij} \mid \mathbf{x})] = \text{mean}_i[E(e_{ij} \mid \mathbf{x})x_{ij}] = \text{mean}_i[e_j x_{ij}] = e_j \text{mean}_i[x_{ij}].$$

The simultaneous equations describing all n samples take the matrix form

$$\boldsymbol{\sigma} = \mathbf{x}\mathbf{e} + \boldsymbol{\varepsilon}, \quad [4]$$

where  $\boldsymbol{\sigma}$  and  $\boldsymbol{\varepsilon}$  are column n-vectors and  $\mathbf{e}$  is a column k-vector. The total scattering  $\boldsymbol{\sigma}$  and species concentrations  $\mathbf{x}$  are fixed by measurement, and the error  $\boldsymbol{\varepsilon}$  carries all of the randomness introduced by variations in the scattering efficiencies. If  $n \geq k$  and  $\boldsymbol{\varepsilon}$  is negligibly small, then [4] provides an overdetermined set of equations for the unknown vector  $\mathbf{e}$  of mean scattering efficiencies, in terms of the measured quantities  $\boldsymbol{\sigma}$  and  $\mathbf{x}$ .

Classical regression analysis estimates  $\mathbf{e}$  as the ordinary least squares solution to [4] (Seber, 1977),

$$\mathbf{e}^{\text{OLS}} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \boldsymbol{\sigma}. \quad [5]$$

The error  $\boldsymbol{\delta} = \mathbf{e}^{\text{OLS}} - \mathbf{e}$  in this estimate depends on the random error in the deterministic relationship:

$$\begin{aligned} \mathbf{e}^{\text{OLS}} &= (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T (\mathbf{x}\mathbf{e} + \boldsymbol{\varepsilon}) \\ &= \mathbf{e} + (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \boldsymbol{\varepsilon}. \end{aligned} \quad [6]$$

Equation [3] implies that  $\boldsymbol{\delta}$  has zero expectation:

$E((\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \boldsymbol{\varepsilon} | \mathbf{x}) = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T E(\boldsymbol{\varepsilon} | \mathbf{x}) = 0$ , because  $E(\boldsymbol{\varepsilon}_i | \mathbf{x}) = E(\sum_j (e_{ij} - e_j) x_{ij} | \mathbf{x}) = \sum_j x_{ij} E(e_{ij} - e_j | \mathbf{x}) = 0$  for each i. Given [3],  $\mathbf{e}^{\text{OLS}}$  is thus an unbiased estimate for  $\mathbf{e}$ :

$$E(\mathbf{e}^{\text{OLS}} | \mathbf{x}) = \mathbf{e}.$$

The error  $\boldsymbol{\delta} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \boldsymbol{\varepsilon}$  in  $\mathbf{e}^{\text{OLS}}$  arises from empirical associations between concentrations and efficiencies. The derivation of [4] shows that  $\boldsymbol{\varepsilon} = \sum_h \boldsymbol{\varepsilon}(h)$ , where  $\boldsymbol{\varepsilon}(h)$  is the vector whose  $i^{\text{th}}$  entry is  $(e_{ih} - e_h) x_{ih}$ ,  $h=1, \dots, k$ . The only optical properties involved in  $\boldsymbol{\varepsilon}(h)$  are those of species h;  $\boldsymbol{\varepsilon}(h) = \mathbf{0}$  for a species with constant efficiency  $e_{ih} \equiv e_h$ . The estimation error is the sum  $\boldsymbol{\delta} = \sum_h \boldsymbol{\delta}(h)$ , where  $\boldsymbol{\delta}(h) = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \boldsymbol{\varepsilon}(h)$ . The only optical properties involved in  $\boldsymbol{\delta}(h)$  are again those of species h;  $\boldsymbol{\delta}_j(h)$  represents the effect of variations  $e_{ih} - e_h$  in the actual efficiency of species h on the estimated efficiency  $e_j^{\text{OLS}}$  of species j. As shown above, fluctuations in efficiencies affect  $\mathbf{e}^{\text{OLS}}$  only when they correlate in some fashion with species mass concentrations; however,  $\boldsymbol{\delta}_j(h)$  is not a simple function of individual correlations.

The error  $\boldsymbol{\delta}_j = \sum_h \boldsymbol{\delta}_j(h)$  in the estimate  $e_j^{\text{OLS}}$  for any species j can be apportioned into contributions associated with concentration dependences in the scattering efficiency of that and each other species h. The aggregate error  $\boldsymbol{\delta}_j = e_j^{\text{OLS}} - e_j$  can also -- and more easily -- be calculated as the difference between the regression estimate and the true mean (Sloane, 1988). The "microscopic" view offered here carries useful additional information, however, even if this

information is not easily interpreted in terms of simple aerosol statistics. The decomposition  $\delta_j = \Sigma_h \delta_j(h)$  can reveal estimates to be accidentally -- and thus ungeneralizably -- successful, because they result from offsetting errors in the estimation model.

## EXAMPLE

This section illustrates the foregoing ideas with a concrete example based on SCAQS measurements. McMurry and Zhang (1991) integrated SCAQS data into various detailed models of the ambient aerosol. We focus here on a model depicting the aerosol as an external mixture of four chemically distinct particle types: nitrate, sulfate, carbonaceous, and soil.

As part of SCAQS, particle size distributions of chemical species were measured with Berner (sulfate and nitrate), MOUDI (carbon), and DRUM (soil iron) impactors in each of 44 sequential observation periods (Zhang et al., 1993). For each species and each period, McMurry and Zhang (1991) computed an individual scattering efficiency  $e_{ij}$ , accounting in the process for water uptakes inferred from TDMA data. The modeled scattering efficiencies thus vary from sample to sample, responding to changes in both size distribution and water content. These variations are evident in the standard deviations of Table 1, which summarizes modeled mass concentrations and scattering efficiencies, and Tables 2 and 3 show they are not necessarily random.

Total fine particle scattering in each sample is the sum of the contributions by individual species,

$$\sigma_i = e_{i,\text{nitrate}}x_{i,\text{nitrate}} + e_{i,\text{sulfate}}x_{i,\text{sulfate}} + e_{i,\text{carbon}}x_{i,\text{carbon}} + e_{i,\text{soil}}x_{i,\text{soil}}.$$

(The concentrations  $x_{i,j}$  appearing here are from the same impactor data used in the calculation of  $e_{i,j}$ ; since the species are modeled as externally mixed, their modeled total scattering is in fact defined by this sum.) The mean over all 44 samples is thus the sum of the mean contributions,

$$\begin{aligned} \text{mean}_i(\sigma_i) = & \text{mean}_i(e_{i,\text{nitrate}}x_{i,\text{nitrate}}) + \text{mean}_i(e_{i,\text{sulfate}}x_{i,\text{sulfate}}) \\ & + \text{mean}_i(e_{i,\text{carbon}}x_{i,\text{carbon}}) + \text{mean}_i(e_{i,\text{soil}}x_{i,\text{soil}}). \end{aligned} \quad [7]$$

Equation [7] provides an exact, causal budget for light scattering, given in column 1 of Table 4.

The mean products on the right-hand side of [7] are typically estimated in practice as the products of the factor means:  $\text{mean}_i(e_{ij}x_{ij}) \approx \text{mean}_i(e_{ij})\text{mean}_i(x_{ij}) = e_j\text{mean}_i(x_{ij})$ . This approach rests on the assumption (equation [3]) that scattering efficiency is statistically independent of concentration. Column 2 of Table 4 gives the components of the resulting approximate budget:

$$\text{mean}_i(\sigma_i) \approx e_{\text{nitrate}}x_{\text{nitrate}} + e_{\text{sulfate}}x_{\text{sulfate}} + e_{\text{carbon}}x_{\text{carbon}} + e_{\text{soil}}x_{\text{soil}}, \quad [8]$$

where  $x_j = \text{mean}_i(x_{ij})$ .

The approximate scattering contributions from [8] agree with the exact contributions from [7] to within 10 percent. Differences arise from empirical associations between scattering efficiency and concentration. The discrepancies can be expressed as follows in terms of sample standard deviations ( $sd_i$ ) and correlation coefficients ( $r_i$ ) over all observations:  $\text{mean}_i(e_{ij}x_{ij}) - \text{mean}_i(e_{ij})\text{mean}_i(x_{ij}) = sd_i(e_{ij})sd_i(x_{ij})r_i(e_{ij},x_{ij})$ . In the case of nitrate, for example, the relevant entries in Tables 1 and 3 yield  $77\text{Mm}^{-1} - 71\text{Mm}^{-1} = (1.5\mu\text{g}/\text{m}^{-3})(8.9\text{m}^2/\text{g})0.44$ .

The mean scattering efficiencies  $e_j$  appearing in [8] are of course usually unknown, and must themselves be estimated. Regression of total scattering on species concentrations yields the estimates  $e_j^{\text{OLS}}$  shown in column 2 of Table 5. Although the regression model fits the data well ( $r=0.960$ ), the true values  $e_{\text{nitrate}}$  and  $e_{\text{soil}}$  (column 1) lie outside the 95% confidence intervals of the regression estimates  $e_{\text{nitrate}}^{\text{OLS}}$  and  $e_{\text{soil}}^{\text{OLS}}$ . Substitution of the regression estimates for the mean scattering efficiencies in [8] yields the regression budget for particle light scattering,

$$\text{mean}_i(\sigma_i) = e_{\text{nitrate}}^{\text{OLS}}x_{\text{nitrate}} + e_{\text{sulfate}}^{\text{OLS}}x_{\text{sulfate}} + e_{\text{carbon}}^{\text{OLS}}x_{\text{carbon}} + e_{\text{soil}}^{\text{OLS}}x_{\text{soil}}, \quad [9]$$

given in column 3 of Table 4. A comparison with the exact budget (column 1) reveals substantial inaccuracies in the regression estimate.

In an ordinary application, a puzzlingly high regression coefficient for nitrates might plausibly be attributed to a negative artifact in the mass measurement. The coefficient would then be interpreted as the ratio of [total] *in situ* nitrate scattering to [partial] filtered nitrate mass, the mass measurement being biased by losses to volatilization. Similarly, an insignificant coefficient for soil might be attributed to imprecise determinations of the trace elements used to estimate soil mass. These explanations are inapplicable here, however, because the nitrate and soil components of the model aerosol supplying our scattering data are precisely as represented by our mass data.

The errors in the regression estimates  $e_j^{\text{OLS}}$ , like those in the approximate budget [8], are traceable to associations in the data between scattering efficiencies and concentrations. Table 6 gives for our example the decomposition of the error vector  $\delta = \sum_h \delta(h)$  derived in the preceding section. Column h is the component vector  $\delta(h)$  arising from concentration dependences in the scattering efficiency of species h; statistically significant effects ( $p \leq 0.05$ ) are underlined. The entry  $\delta_j(h)$  on row j represents the error these contribute to the estimate for species j. The sum of all entries in row j is the total error  $e_j^{\text{OLS}} - e_j = \delta_j$ ; for example,  $e_{\text{nitrate}}^{\text{OLS}} - e_{\text{nitrate}} = 7.5 - 5.4 = 1.3 + 1.0 - 0.3 + 0.1$ . Statistical significance was determined by bootstrap resampling.

Table 6 illustrates differing ways in which regression analysis can yield unreliable

estimates of scattering efficiency. The significant positive error  $e_{\text{nitrate}}^{\text{OLS}} - e_{\text{nitrate}}$  reflects the reinforcing effects of significant concentration dependences in two species' scattering efficiencies. The negative error  $e_{\text{soil}}^{\text{OLS}} - e_{\text{soil}}$ , whose statistical significance is less pronounced, reflects the reinforcing effects of individually insignificant concentration dependences in all four species' scattering efficiencies. Only the sulfate estimate  $e_{\text{sulfate}}^{\text{OLS}}$  is "right for the right reasons," involving no significant violations of the statistical assumption [3].

The most troubling result concerns the carbon estimate, which seems almost as accurate as the sulfate estimate. The regression coefficient  $e_{\text{carbon}}^{\text{OLS}}$  is well within its estimated uncertainty ( $p \leq 0.05$ ) of the true value. Our microscopic view suggests this superficial agreement is fortuitous, however. Table 6 shows that it conceals offsetting errors produced by significant concentration dependences in  $e_{\text{sulfate}}$  and  $e_{\text{carbon}}$ .

Our analysis to this point has not explicitly accounted for the effects of water uptake by hygroscopic species. The high correlation observed in Table 3 between the scattering efficiencies of nitrate and sulfate suggests that both are driven by ambient variations in relative humidity. The influence of humidity on nitrate and sulfate scattering can be addressed by the conventional transformations  $e_{ij}x_{ij} = [e_{ij,\text{dry}}/(1-\text{RH}_i)]x_{ij} = e_{ij,\text{dry}}x_{ij,\text{wet}}$  described earlier. The often-disregarded influence of humidity on organic carbon scattering also deserves attention (Saxena and Hildemann, 1996), but is as yet poorly characterized. For simplicity and consistency with past practice (e.g. Groblicki et al., 1981),  $x_{\text{carbon,wet}}$  and  $x_{\text{soil,wet}}$  are accordingly set equal to  $x_{\text{carbon}}$  and  $x_{\text{soil}}$  in what follows.

Regression of total scattering on the RH-adjusted concentrations  $x_{i,\text{nitrate}}/(1-\text{RH}_i)$ ,  $x_{i,\text{sulfate}}/(1-\text{RH}_i)$ ,  $x_{i,\text{carbon}}$ , and  $x_{i,\text{soil}}$  yields the estimates shown in column 4 of Table 5. Note that the transformation of nitrate and sulfate concentrations changes the regression coefficients for the untransformed carbon and soil concentrations. Table 7 shows the structure of the errors also to be quite different. Substitution of the new estimates in [8] yields a revised scattering budget, 
$$\text{mean}_i(\sigma_i) = e_{\text{nitrate,dry}}^{\text{OLS}}x_{\text{nitrate,wet}} + e_{\text{sulfate,dry}}^{\text{OLS}}x_{\text{sulfate,wet}} + e_{\text{carbon}}^{\text{OLS}}x_{\text{carbon}} + e_{\text{soil}}^{\text{OLS}}x_{\text{soil}}, \quad [10]$$
 given in column 4 of Table 4.

Accounting for RH in the regression yields mixed results for our example. The new estimates of scattering efficiencies for nitrate and soil are within their estimated uncertainties ( $p \leq 0.05$ ) of the true values (Table 5), and yield much more realistic contributions to scattering (Table 4). The RH-adjusted sulfate efficiency remains close to the true RH-adjusted value (Table 5), but yields a less realistic contribution to scattering (Table 4) when multiplied by RH-adjusted concentration. The fortuitous accuracy of the carbon estimate is lost, with the offsetting errors of

the earlier version replaced by reinforcing errors in the new (Table 7).

Figure 1 summarizes the foregoing estimates of scattering efficiency, along with others from regressions allowing the intercept to float, a common option. For the particular set of observations examined here, the chemically resolved scattering efficiencies derived from regression analysis are not obviously more informative than the overall mean ratio of total scattering to total mass.

Before leaving this example, we may note that our model data set was an ideal candidate for regression analysis.

- i The functional dependence of total scattering on species concentrations is exactly linear, because the species are externally mixed.
- ii The regression model is perfectly specified, because there are no “unmeasured” species, no “background” scattering.
- iii The predictor variables are known without error, because the response variable is calculated from these as measured.

The errors in our apportionments resulted from the inevitable tradeoff for regression’s limited data requirements. The non-random variations in scattering efficiency which biased our estimates were simply undetectable, without information that went beyond species concentrations and total scattering.

## NON-EXTERNAL MIXTURES

Detailed analyses by McMurry and Zhang (1991) indicated that some species in the actual SCAQS aerosol were probably mixed within individual particles. This section briefly indicates the methodological adjustments that must be made to accommodate more realistic aerosols in which distinct species are not distinct sub-collections of particles.

The fundamental complication introduced by non-external mixtures is that individual species no longer possess unambiguous scattering coefficients of their own (White, 1986). Unlike scattering by separate particles, the scattering by a single particle is not a sum of contributions from its constituent parts. Various accounting schemes can be used to construct within-particle scattering budgets, but the resulting apportionments are necessarily somewhat conventional. Their attribution of scattering to a species does not necessarily yield the quantity of practical interest, which is the scattering decrement to be expected from the reduction or elimination of that species.

The problem is easily seen if we imagine two species of Rayleigh scatterers, of similar size and refractive index. A mixed dimer particle, formed from one monomer of each species, scatters approximately four times the energy either monomer does on its own. Thus, although neither component “contributes” more scattering than the other, removing either monomer reduces dimer scattering by more than half.

Scattering efficiency can still be defined as the ratio of scattering decrement to mass decrement when a species is removed in a specified manner, the decrements referring to the (observable) properties of the whole aerosol (White, 1986). This generalizes our previous usage for external mixtures, because the species contributions in that case do add to the aerosol total. As shown by our Rayleigh illustration, however, the correspondingly generalized “contributions”  $e_j x_j$  need no longer add to the total  $\sigma_{sp}$ .

Our Rayleigh illustration shows that the accounting described by equation [1] need not balance when species are non-externally mixed. Regression analysis can regard imbalances in individual observations as errors of either model formulation or extinction measurement. If we expect the imbalances to be small and vary randomly from one observation to the next, we may proceed as before to estimate mean scattering efficiencies as the regression coefficients in [2]. If we expect the imbalances in [1] to favor one side consistently over the other, we may instead allow the intercept in our regression to float. In either approach the general analysis of estimation error must expand to include terms involving correlations with the imbalance.

Figure 2 shows exact and estimated scattering efficiencies for a model of the SCAQS aerosol as an internal mixture. Although aerosol mixing structure is critical to the theory, it has little practical effect on scattering efficiencies modeled or estimated for the SCAQS observations.

## SUMMARY

The foregoing sections examined the theoretical basis for apportioning light scattering among species whose particle size distribution and water uptake can vary from sample to sample. For clarity and simplicity, we focused on the ideal case of external mixtures and accurate measurements. We demonstrated in this setting that multiple linear regression yields unbiased apportionments, under a statistical condition that may or may not be satisfied in the actual atmosphere.

The theory developed here requires that each species' scattering efficiency vary independently of all species' mass concentrations. Explicit formulas can be derived that relate

violations of this condition to biases in the regression estimates. These formulas allow us to resolve the aggregate errors in regression estimates into distinct components, each traceable to a different species' scattering efficiency. The decompositions can show "accidentally" accurate estimates to conceal offsetting component errors. This line of analysis was illustrated with a concrete example based on actual measurements of species size distributions and water uptake.

This paper identifies theoretical conditions that justify regression apportionments of light scattering, but does not address the generality with which these conditions are satisfied in actual applications. Application of our methodological tools to a variety of other settings is needed to establish the representativeness of our empirical results.

#### ACKNOWLEDGMENT

This research benefitted from the encouragement of Pradeep Saxena, project manager of supporting EPRI contracts with Washington University and the University of Minnesota.

#### REFERENCES

- Anderson T.L., Charlson R.J., White W.H., and McMurry P.H. (1994) Comment on "Light scattering and cloud condensation nucleus activity of sulfate aerosol measured over the Northeast Atlantic Ocean" by D.A. Hegg et al. *Journal of Geophysical Research* 99, 25947-25949.
- Appel B.R., Tokiwa Y., Hsu J., Kothny E.L., and Hahn E. (1985) Visibility as related to atmospheric aerosol constituents. *Atmospheric Environment* 19, 1525-1534.
- Bohren C.F. and Huffman D.R. (1983) *Absorption and Scattering of Light by Small Particles*. John Wiley & Sons, New York.
- Cass G.R. (1979) On the relationship between sulfate air quality and visibility with examples in Los Angeles. *Atmospheric Environment* 13, 1069-1084.
- Charlson R.J., Covert D.S., Larson T.V., and Waggoner A.P. (1978) Chemical properties of tropospheric sulfur aerosols. *Atmospheric Environment* 12, 39-53.
- Groblicki P.J., Wolff G.T., and Countess R.J. (1981) Visibility-reducing species in the Denver "brown cloud" -- I. Relationships between extinction and chemical composition. *Atmospheric Environment* 15, 2473-2484.

- Jaenicke R. (1978) Physical properties of atmospheric particulate sulfur compounds. *Atmospheric Environment* 12, 161-169.
- Lowenthal D.H., Rogers C.F., Saxena P., Watson J.G., and Chow J.C. (1995) Sensitivity of estimated light extinction coefficients to model assumptions and measurement errors. *Atmospheric Environment* 29, 751-766.
- McMurry P.H. and Stolzenburg M.R. (1988) On the sensitivity of particle size to relative humidity for Los Angeles aerosols. *Atmospheric Environment* 23, 497-507.
- McMurry P.H. and Zhang X. (1991) *Optical Properties of Los Angeles Aerosols: an Analysis of Data Acquired During SCAQS*. Particle Technology Laboratory Report No. 775, University of Minnesota, Minneapolis.
- McMurry P.H., Zhang X., and Lee C.T. (1996) Issues in aerosol measurement for optics assessments. *Journal of Geophysical Research* 101, 19189-19197.
- Saxena P. and Hildemann L.M. (1996) Water-soluble organics in atmospheric particles: a critical review of the literature and application of thermodynamics to identify candidate compounds. *Journal of Atmospheric Chemistry* 24, 57-109.
- Seber G.A.F. (1977) *Linear Regression Analysis*. John Wiley & Sons, New York.
- Sisler J.F. and Malm W.C. (1994) The relative importance of soluble aerosols to spatial and seasonal trends of impaired visibility in the United States. *Atmospheric Environment* 28, 851-862.
- Sloane C.S. (1983) Optical properties of aerosols -- comparison of measurements with model calculations. *Atmospheric Environment* 17, 409-416.
- Sloane C.S. (1988) Forecasting visibility impairment: a test of regression estimates. *Atmospheric Environment* 22, 2033-2045.
- Trijonis J. (1979) Visibility in the Southwest -- an exploration of the historical data base. *Atmospheric Environment* 13, 833-843.

- Vasconcelos L.A. de P., Macias E.S., and White W.H. (1994) Aerosol composition as a function of haze and humidity levels in the southwestern U.S. *Atmospheric Environment* 28, 3679-3691.
- Vasconcelos L.A. de P., Macias E.S., McMurry P.H., Turpin B.J., and White W.H. (1998) . The physical significance of regressions relating light scattering to aerosol composition. Manuscript in preparation.
- White W.H. (1976) Reduction of visibility by sulphates in photochemical smog. *Nature* 264, 735-736.
- White W.H. (1986) On the theoretical and empirical basis for apportioning extinction by aerosols: a critical review. *Atmospheric Environment* 20, 1659-1672.
- White W.H. and Macias E.S. (1987) On measurement error and the empirical relationship of atmospheric extinction to aerosol composition in the non-urban west. *APCA Transactions* 10, 783-794.
- White W.H., Macias E.S., Nininger R.C., and Schorran D. (1994) Size-resolved measurements of light scattering by ambient particles in the southwestern U.S.A. *Atmospheric Environment* 28, 909-921.
- Zhang X.Q., McMurry P.H., Hering S.V., and Casuccio G.S. (1993) Mixing characteristics and water content of submicron aerosols measured in Los Angeles and at the Grand Canyon. *Atmospheric Environment* 27A, 1593-1607.
- Zhang X., Turpin B.J., McMurry P.H., Hering S.V., and Stolzenburg M.R. (1994) Mie theory evaluation of species contributions to 1990 wintertime visibility reduction in the Grand Canyon. *Journal of the Air & Waste Management Association* 44, 153-162.

j	mean(x <sub>j</sub> ) μg/m <sup>3</sup>	sd(x <sub>j</sub> ) μg/m <sup>3</sup>	mean(e <sub>j</sub> ) m <sup>2</sup> /g	sd(e <sub>j</sub> ) m <sup>2</sup> /g
Nitrate	13.3	8.9	5.4	1.5
Sulfate	9.5	4.6	5.1	1.4
Carbon	12.3	5.8	5.8	0.7
Soil	4.8	2.4	2.8	1.5

**Table 1.** Summary statistics for external mixture model of SCAQS aerosol. Columns give arithmetic means and standard deviations of species mass concentrations and scattering efficiencies.

correlation, r		x <sub>j</sub>			
		Nitrate	Sulfate	Carbon	Soil
e <sub>j</sub>	Nitrate	1	0.47	0.43	+
	Sulfate	0.93	1	0.38	-
	Carbon	+	+	1	+
	Soil	-	-	-	1

**Table 2.** Pearson coefficients for inter-species correlations of mass concentration (above diagonal) and scattering efficiency (below diagonal) in external mixture model of SCAQS aerosol. Signs are indicated for all correlations; values are given only where statistically significant ( $p \leq 0.05$ ).

correlation, r		x <sub>j</sub>			
		Nitrate	Sulfate	Carbon	Soil
e <sub>j</sub>	Nitrate	0.44	+	-	-
	Sulfate	0.44	+	-	-.39
	Carbon	-	+	0.34	-
	Soil	+	-	-	-

**Table 3.** Pearson coefficients for correlation of scattering efficiency (e<sub>j</sub>) with mass concentration (x<sub>j</sub>) in external mixture model of SCAQS aerosol. Signs are indicated for all correlations; values are given only where statistically significant ( $p \leq 0.05$ ).

j	$m(e_j x_j)$ equation [7]	$m(e_j)m(x_j)$ equation [8]	$e_j^{\text{OLS}}m(x_j)$ equation [9]	$e_{j,\text{dry}}^{\text{OLS}}m(x_{j,\text{wet}})$ equation [10]
Nitrate	77	71	100	69
Sulfate	50	48	55	38
Carbon	73	72	58	97
Soil	13	13	- 1	11

**Table 4.** Exact (first column) and estimated (second through fourth columns) contributions ( $\text{Mm}^{-1}$ ) to mean scattering in external mixture model of SCAQS aerosol. Column headings are simplified from notation in text:  $m(x_j) = \text{mean}_i(x_{ij})$ ,  $m(e_j) = \text{mean}_i(e_{ij})$ , etc.

j	$\text{mean}(e_j)$ $\text{m}^2/\text{g}$	$e_j^{\text{OLS}}$ $\text{m}^2/\text{g}$	$\text{mean}(e_{j,\text{dry}})$ $\text{m}^2/\text{g}$	$e_{j,\text{dry}}^{\text{OLS}}$ $\text{m}^2/\text{g}$
Nitrate	5.4	<u>7.5±0.6</u>		
Sulfate	5.1	5.8±1.0		
Carbon	5.8	4.7±0.8	5.8	<u>7.9±0.6</u>
Soil	2.8	<u>-.3±1.5</u>	2.8	2.3±1.4
Nitrate/(1-RH)			2.2	1.9±0.2
Sulfate/(1-RH)			2.1	1.6±0.4

**Table 5.** Mean scattering efficiencies for external-mixture model of SCAQS aerosol, compared with regression estimates from equation [5]. Regression coefficients are accompanied by standard errors, and are underlined if they disagree ( $p \leq 0.05$ ) with the true means.

$\delta_j(h), m^2/g$	$\delta(h)$				
		Nitrate	Sulfate	Carbon	Soil
	Nitrate	<u>1.3</u>	<u>1.0</u>	-.3	0.1
j	Sulfate	0.4	0.2	0.2	-.1
	Carbon	-.9	<u>-.7</u>	<u>0.5</u>	0.1
	Soil	-1.2	-1.0	-.6	-.3

**Table 6.** Decomposition of estimation errors  $e_j^{OLS} - e_j$  in Table 5. Entries in column h arise from concentration-dependent variations in the scattering efficiency  $e_h$  of species h; the sum of the entries in row j is the error  $e_j^{OLS} - e_j$  in the regression estimate for species j. Statistically significant ( $p \leq 0.05$ ) values are underlined.

$\delta_j(h), m^2/g$	$\delta(h)$				
		Nitrate	Sulfate	Carbon	Soil
	Nitrate	-.1	-.1	-.1	0.0
j	Sulfate	-.3	-.2	0.1	-.1
	Carbon	0.7	<u>0.7</u>	<u>0.5</u>	0.1
	Soil	0.5	-.2	-.6	-.2

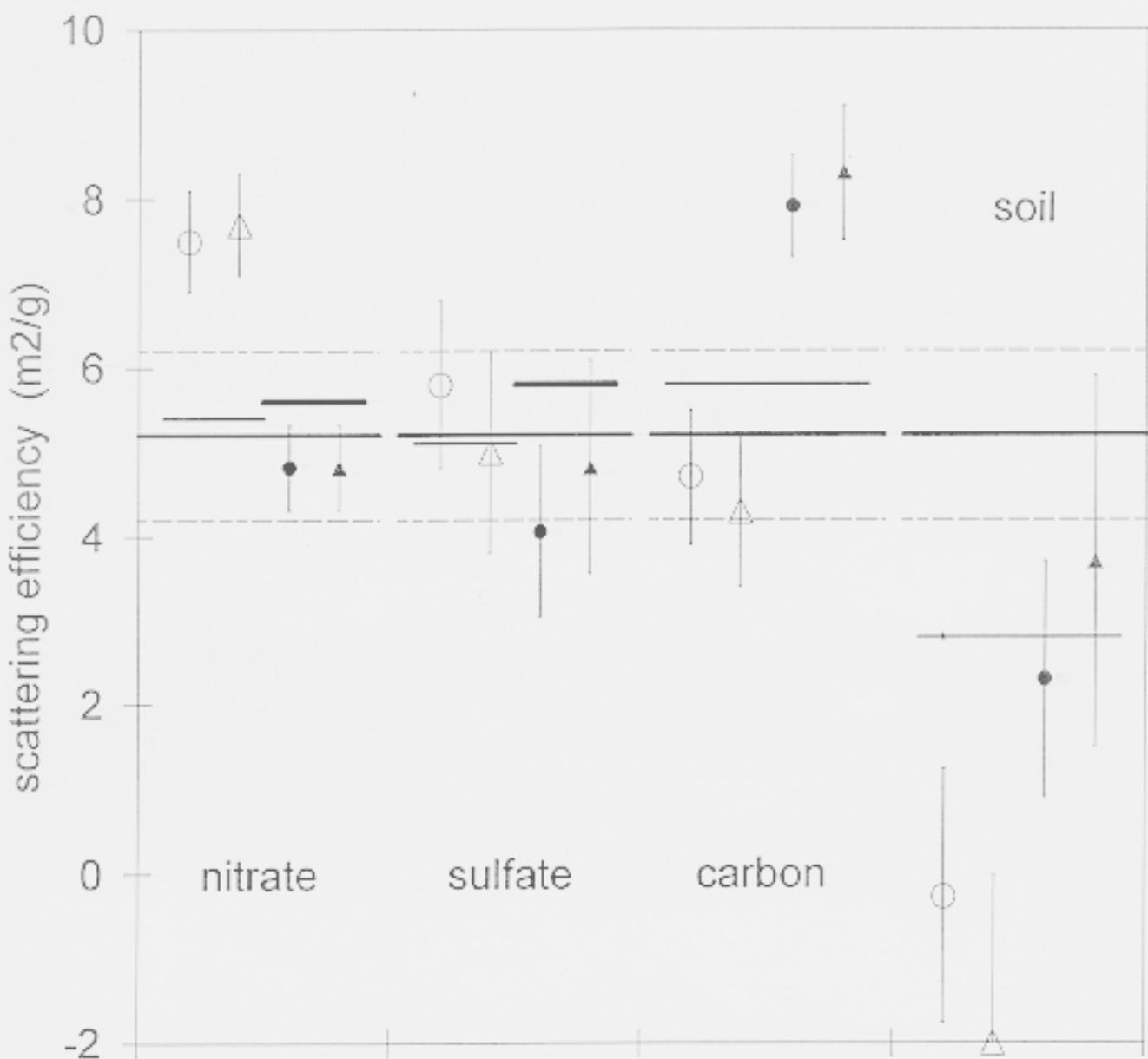
**Table 7.** Decomposition of estimation errors  $e_{j,dry}^{OLS} - e_{j,dry}$  in Table 5. Entries in column h arise from concentration-dependent variations in the dry scattering efficiency  $e_{h,dry}$  of species h; the sum of the entries in row j is the error  $e_{j,dry}^{OLS} - e_{j,dry}$  in the regression estimate for species j. Statistically significant ( $p \leq 0.05$ ) values are underlined.

**Figure 1.** Mean scattering efficiencies for the external mixture model of the SCAQS aerosol. Short horizontal lines indicate exact model values for the indicated chemical species. Symbols indicate estimates from regression with zero or floating intercept, and implicit or explicit treatment of humidity. For the implicit RH treatment, the exact  $\text{mean}(e_j)$  and zero-intercept  $e_j^{\text{OLS}}$  are from respectively the first and second columns of Table 5. For the explicit RH treatment, the corresponding  $\text{mean}(e_{j,\text{dry}})$  and  $e_{j,\text{dry}}^{\text{OLS}}$  from the third and fourth columns of Table 5 are multiplied by  $\text{mean}(1/(1-\text{RH}))$  to facilitate visual comparison.

The long horizontal line marks the observable mean efficiency of the total aerosol. Standard deviation and standard errors are indicated for the bulk efficiency and regression estimates.

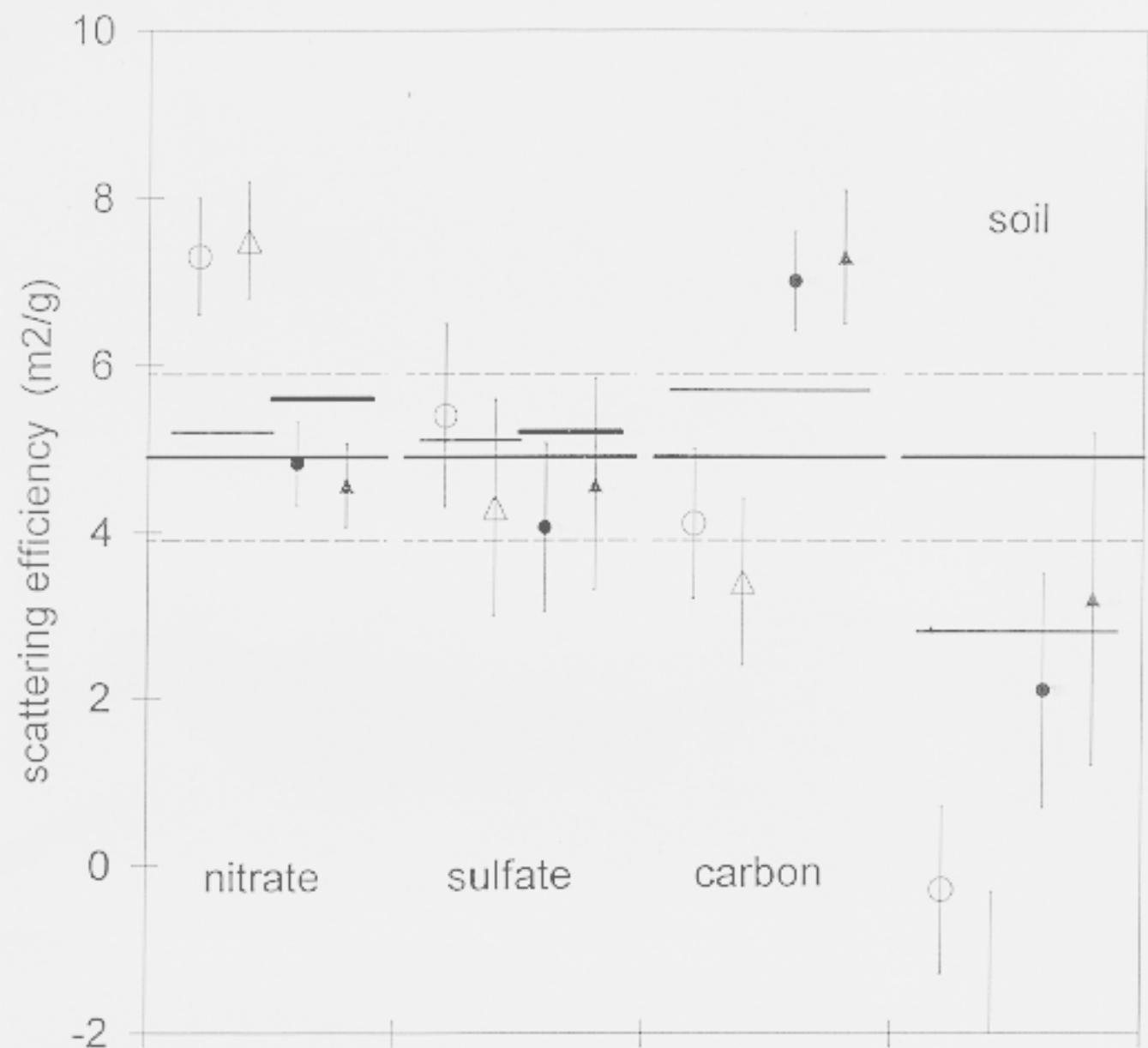
**Figure 2.** Mean scattering efficiencies for the internal mixture model of the SCAQS aerosol. Plotting conventions are as in Figure 1.

# SCAQS - external



- zero intercept, implicit RH    △ float intercept, implicit RH
- zero intercept, explicit RH    ▲ float intercept, explicit RH

# SCAQS - internal



○ zero intercept, implicit RH    △ float intercept, implicit RH

● zero intercept, explicit RH    ▲ float intercept, explicit RH

