APPENDIX 6B: Tracer Mass Balance Model Regression (TMBR) Model
and
Tracer Mass Balance (TMB) Model

Overview

The Tracer Mass Balance Regression Model is a multiple regression based model which may be used to apportion an aerosol species of interest measured at a receptor site to the various contributing sources. It has been shown to be a special case of the General Mass Balance (GMB) Model. The actual regression analysis may be performed using the method of least squares. However, since the independent variables in this model are ambient concentrations of various aerosol components which are measured with error, the method of Orthogonal Distance Regression (ODR) is expected to give better estimates of the source contributions. A detailed discussion of the method of ODR may be found in the book by Fuller(1987).

Model Equations

The basic equation for TMBR model equation is:

$$C_{ik} = \gamma_0 + \sum_{u=1}^{h} \gamma_{i_u} C_{i_u k} \phi_{uk} \qquad (1)$$

where:

C_{ik} = concentration of species i at the receptor for time period k. In the current application i refers to Sulfate Sulfur or SO2 sulfur.

$C_{i_u k}$ = concentration of trace element i_u which serves as a tracer for a group of one or more sources, for time period k.

γ_{i_u} = regression coefficient for trace element i_u which acts as a tracer for a group of one or more sources.

γ_0 = intercept representing the mean background concentration of the species of interest, at the receptor.

h = number of groups of sources, each group being represented by a particular aerosol species which acts as a tracer for that group of sources.

ϕ_{uk} = a factor which is a function of field measurements, sampling period and possibly source type, chosen in such a way that the γ coefficients in the model (1) are invariant with respect to the sampling period.

The model is known as the tracer mass balance (TMB) model when only a single trace element is used as a tracer for a particular source and all the remaining sources are accounted for by the intercept term in the model. When several trace elements are used in addition to the tracer for the distinguished source of interest, then the model is referred to as tracer mass balance regression (TMBR) model. The simplest versions of the TMBR model and the TMB model use $\phi_{uk} = 1$ for all time periods and source groups. In the current application we have used $\phi_{uk} = 1$ as well as $\phi_{uk} = RH_k$ where RH_k is the relative humidity at the receptor during sampling period k.

The use of RH_k as a linear factor in the above model was motivated by the following considerations. In apportioning a secondary aerosol, the constant $\beta_{i_u jk}$ derived from the GMB model had the form

$$\beta_{i_u jk} = \frac{r^*_{1jk}}{r_{i_u jk}} \frac{c_{2jk}}{c_{i_u jk}} \tag{2}$$

with

$$r^*_{ijk} = \frac{K_c(i^*, j, k)}{K_c(i^*, j, k) + K_d(i^*, j, k) - K_d(i, j, k)} \times \{exp(-K_d(i, j, k)t_{jk}) - exp(-[K_c(i^*, j, k) + K_d(i^*, j, k)]t_{jk})\} \tag{3}$$

and

$$r_{i_u jk} = exp(-(K_c(i_u, j, k) + K_d(i_u, j, k))t_{jk}) \tag{4}$$

If the species i_u does not convert and its deposition rate is the same as that of the secondary aerosol species i being apportioned, then

$$r_{i_u jk} = exp(-K_d(i, j, k)t_{jk}) \tag{5}$$

so that the ratio $r^*_{1jk}/r_{i_u jk}$ reduces to $K_c(i^*, j, k)t_{jk}$ after using the approximation

$$exp(x) \approx 1 + x \text{ (when } x \text{ is sufficiently small).} \tag{6}$$

Recall that the full infinite series expansion for $exp(x)$ is given by

$$exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

and we have used a first order approximation in (6). It is possible to use higher order approximations of $exp(x)$ in these derivations but this is not pursued here.

Assuming that $K_c(i^*, j, k)$ is proportional to RH_k with proportionality constant B, we obtain that the ratio r^*_{1jk}/r_{i_ujk} is equal to $Bt_{jk}RH_k$ which gives

$$\beta_{i_ujk} = Bt_{jk}RH_k\frac{c_{2jk}}{c_{i_ujk}}. \tag{7}$$

Defining

$$\gamma_{i_uk} = \beta_{i_uk}/\phi_{uk} \tag{8}$$

where $\phi_{uk} = RH_k$ and assuming that γ_{i_uk} are constant for all sampling periods rather than the quantities β_{i_uk} suggests the use of RH_k as a linear factor in the TMBR model equation (1).

For purposes of attributing total sulfur or sulfate sulfur to NGS, SCD_4 is a unique NGS tracer. Furthermore, As was found to be below the detectible limit in samples gathered from within the NGS pluem (refer to Table ??). Therefore, As is considered to be a unique tracer for emissions other than NGS and most probably associated with copper smelter emissions.

Therefore, in actual application of the TMBR model to WHITEX data, we have grouped the sources into 3 categories:

- NGS with CD4 serving as the tracer
- Sources with Arsenic (As) serving as a tracer, and,
- all remaining sources, if any.

In the application of the TMB model, there are only two categories, viz, NGS with CD4 as a tracer and all remaining sources.

The TMBR Model equations used in the application are:

$$C_k = \gamma_0 + \sum_{u=1}^{h} \gamma_{i_u} C_{i_uk}\phi_{uk} \tag{9}$$

where:

C_k = concentration of sulfate sulfur or total sulfur for time period k

C_{i_uk} = concentration of trace element i_u for time period k

γ_{i_u} = regression coefficient associated with trace element i_u

γ_0 = intercept representing the mean background concentration of the species being apportioned, due to all sources not accounted for explicitly.

$\phi_{uk} = RH_k$, the relative humidity at the receptor during sampling period k , or 1, depending on the particular application.

All of the cases considered may be written in the form

$$C_k = \gamma_0 + \sum_{u=1}^{2} \gamma_{i_u} A_{uk} \tag{10}$$

where:

A_{uk} = concentration of trace element i_u for time period k or concentration of trace element i_u multiplied by RH_k.

The model is known as the tracer mass balance (TMB) model when the only trace element used is CD4 or scaled CD4 (SCD4). When other trace elements are used in addition to CD_4 then the model is referred to as tracer mass balance regression (TMBR) model. Multiplication by RH, when included, is a surrogate for the RH dependent oxidation rate of SO2 to SO4.

Model Calculations and Uncertainties

The concentrations C_{uk} of sulfate sulfur or total sulfur associated with each trace element i_u for each time period are calculated by multiplying the measured values of A_{uk} for each trace element by the respective regression coefficients as follows.

C_{0k} would just be the intercept representing the contribution from all sources not explicitly accounted for by any of the reference species used in the TMBR model.

$$C_{uk} = \gamma_{i_u} \times A_{uk} \tag{11}$$

The uncertainties for each of these concentrations are calculated by:

$$\sigma_{C_{uk}} = \sqrt{A_{uk}^2 \sigma_{\gamma_{i_u}}^2 + \gamma_{i_u}^2 \sigma_{A_{uk}}^2 + \sigma_{\gamma_{i_u}}^2 \sigma_{A_{uk}}^2} \tag{12}$$

The quantities $\sigma_{A_{uk}}$ are part of the WHITEX data base. The quantities γ_{i_u} are obtained as outputs from the regression packages that were used. Errors in A_{uk} and the estimated regression coefficients have been assumed to be uncorrelated.

The total calculated sulfur C_k for each time period is the sum of the C_{uk}'s summed over all the reference aerosol species i_u and the intercept.

$$C_k = C_0 + \sum_{u=1}^{h} C_{uk} \tag{13}$$

The uncertainty associated with the total calculated sulfur concentration for each time period is:

$$\sigma_{C_k} = \sqrt{\sigma_{C_0}^2 + \sum_{u=1}^{h} \sigma_{C_{uk}}^2} \tag{14}$$

assuming the covariance terms arising in the derivation are negligible. The sources assumed to be associated with each trace element are:

- CD4 or SCD4 – Navajo Generating Station (NGS)
- Selenium (Se) – All power plants including NGS
- Arsenic (As) – Copper smelters

- Intercept – Mean background concentration.

The estimated fraction of sulfur from each source for any given time period is equal to the sulfur associated with the trace element divided by the total calculated sulfur concentration:

$$F_{uk} = \frac{C_{uk}}{C_k} \tag{15}$$

The uncertainty for each of these fractions is:

$$\sigma_{F_{uk}} = \sqrt{\frac{\sigma^2_{C_{uk}}}{C^2_k} + \frac{C^2_{jk}\sigma^2_{C_k}}{C^4_k}} \tag{16}$$

The mean fraction $\overline{F}_u$ of the sulfur attributed to each source is estimated by the mean sulfur concentration $\overline{C}_u$ for that source divided by the mean total calculated sulfur $\overline{C}$.

$$\overline{F}_u = \frac{\overline{C}_u}{\overline{C}} \tag{17}$$

where

$$\overline{C}_u = \frac{1}{s}\sum_{k=1}^{s} C_{uk} \tag{18}$$

and

$$\overline{C} = \frac{1}{s}\sum_{k=1}^{s} C_k. \tag{19}$$

The uncertainties for $\overline{C}_u$ and $\overline{C}$ are calculated by:

$$\sigma_{\overline{C}_u} = \frac{1}{K}\sqrt{\sum_{k=1}^{s} \sigma^2_{C_{jk}}} \tag{20}$$

and

$$\sigma_{\overline{C}} = \frac{1}{K}\sqrt{\sum_{k=1}^{s} \sigma^2_{C_k}}. \tag{21}$$

The uncertainties associated with the mean fractions are calculated by

$$\sigma_{\overline{F}_u} = \sqrt{\frac{\sigma^2_{\overline{C}_u}}{\overline{C}^2} + \frac{\overline{C}^2_k \sigma^2_{\overline{C}}}{\overline{C}^4}}. \tag{22}$$

The uncertainty formulas are all derived using propagation of error methods and assuming the covariances between various terms occurring in the derivation are negligible.

Model Assumptions.

The regression coefficients, including the intercept term, in the model have been assumed to be time independent. The aerosol species used in the model are assumed to be tracers for nonoverlapping groups of sources. In particular, none of the species other than the tracer associated with the source of interest can be emitted by that source unless there is an independent method such as CMB modeling to partition the ambient species concentrations into components attributable to the various groups of sources.

Potential Deviations from Assumptions.

It is highly unlikely that the regression coefficients are constant for all sampling periods. This will inflate the uncertainty in the final apportionments but the extent to which this inflation occurs will depend on how variable the regression coefficients are. We investigate below the possible effects of nonconstant regression coefficients in the TMB model. A similar investigation may be carried out for the more general TMBR model but the derivations are rather cumbersome and details are omitted here. For reasons of convenience, the notation in the subsequent subsection is entirely independent of the rest of the appendix but this need not cause any confusion.

Effect of nonconstant regression coefficients in the TMB model.

Suppose

y_t = pollutant concentration at the receptor at time t.
x_t = concentration of tracer at the receptor at time t.
w_t = pollutant concentration at the receptor attributable to the source under study.
z_t = pollutant concentration at the receptor attributable to other sources.

Then

$$y_t = w_t + z_t. \tag{23}$$

We define

$$m_t = w_t/x_t \tag{24}$$

so that

$$y_t = m_t x_t + z_t \tag{25}$$

It may be desirable to account for the fact that the actual measurements of $\{y_t\}$, $\{x_t\}$ involve measurement errors. Suppose the observed quantities are $\{Y_t\}$, $\{X_t\}$ where

$$\begin{aligned} Y_t &= y_t + S_t \\ X_t &= x_t + E_t\,, \end{aligned} \tag{26}$$

$\{S_t\}$, $\{E_t\}$ being the independent set of measurement errors with means equal to 0 and known standard deviations equal to σ_S, σ_E, respectively. An estimate of the average contribution of the

pollutant by the source under study is given by

$$\overline{\mathrm{NGS}} = \hat{\beta}\bar{X}, \tag{27}$$

where $\hat{\beta}$ is the slope of a structural regression line fit obtained by regressing $\{Y_t\}$ on $\{X_t\}$, while the estimated average fractional contribution, $\bar{f}$, of the source to the receptor, for the duration of the study, is

$$\bar{f} = \frac{\hat{\beta}\bar{X}}{\bar{Y}}. \tag{28}$$

We now investigate how the estimated average pollutant concentrations due to NGS can differ from the actual value for the time period in question.

In the following discussion, a quantity such as $\beta\{y, x\}$ will refer to the slope of the least squares line fitted to $\{(y_t, x_t) \mid t = 1, ..., n\}$, with $\{y_t\}$ as observations on a dependent variable and $\{x_t\}$ as observations on an independent variable. A quantity such as $\bar{x}$ will represent $\frac{1}{n}\sum_{t=1}^{n} x_t$ and σ_x^2 will represent $\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2$.

The true average contribution of the pollutant from NGS to the receptor site is $\bar{w} = \frac{1}{n}\Sigma w_t$. The estimated average contribution is $\hat{\beta}\bar{x}$, where $\hat{\beta}$ is the slope of the regression line fitted to the data $\{(Y_t, X_t) \mid t = 1, ..., n\}$. At first we will consider the situation when $\hat{\beta}$ is the least squares estimate in which case we write $\hat{\beta}_{LS}$.

It is easily verified that

$$\begin{aligned} \Delta &= \hat{\beta}_{LS}\bar{X} - \bar{w} \\ &= \frac{(\bar{x} + \bar{E})(\beta\{w, x\} + \beta\{z, x\} + \beta\{S, x\} + \lambda\beta\{w, E\} + \lambda\beta\{z, E\} + \lambda\beta\{S, E\})}{1 + 2\beta\{E, x\} + \lambda} - \bar{w} \end{aligned} \tag{29}$$

where $\lambda = \sigma_E^2/\sigma_x^2$ and Δ is the difference between the estimated average NGS contribution and the true average NGS contribution. It seems reasonable to assume that the quantities

$$\bar{E}, \beta\{S, x\}, \beta\{w, E\}, \beta\{z, E\}, \beta\{S, E\}, \beta\{E, x\} \tag{30}$$

are all nearly zero because we expect measurement errors E_t averaged over n time periods to be nearly zero and because we expect measurement errors to be uncorrelated with the true values x, z and w.

To this degree of approximation,

$$\Delta \approx \frac{\bar{x}\beta\{w, x\} - \bar{w} + \bar{x}\beta\{z, x\} - \lambda\bar{w}}{1 + \lambda}. \tag{31}$$

If $\hat{\beta}$ is the estimate obtained using structural regression (or Orthogonal Distance Regression (ODR)), denoted by $\hat{\beta}_{ODR}$, we would obtain

$$\Delta \approx (\bar{x}\beta\{w, x\} - \bar{w}) + \bar{x}\beta\{z, x\} \tag{32}$$

since $\hat{\beta}_{ODR} \approx \hat{\beta}_{LS}(1 + \lambda)$, where Δ is the difference between estimated and actual average NGS contribution during the time period under study.

The quantity $\bar{x}\beta\{w,x\} - \bar{w}$ is zero if w_t/x_t is constant and will differ from zero if the least squares line fitted to the points $\{(w_t, x_t) \mid t = 1, ..., n\}$ has a nonzero intercept. On the other hand, the quantity $\beta\{z,x\}$ is zero or nonzero depending on whether the least squares line fitted to the points $\{(z_t, x_t) \mid t = 1, ..., n\}$ has a zero slope or not, i.e., whether or not z_t and x_t are "correlated". Ideally, if there was a constant background pollutant concentration $z_t \equiv \bar{z}$ and if the tracer release was directly proportional to emissions, and emissions were conservative, so that $m_t \equiv \bar{m}$, we would have $\Delta \approx 0$ and the reported estimated average NGS contribution should be a reliable estimate of the actual value for the time period in question.

Model Inputs.

The model requires the following quantities as inputs:

- The ambient concentrations of the aerosol species being apportioned, which is SO4 in our application.
- The ambient concentrations of the reference or tracer species, CD4 and As.
- Relative humidity at the receptor for each of the sampling periods, when $\phi_{uk} = RH_k$ is used in the model rather than $\phi_{uk} = 1$.
- The uncertainties in the above quantities, when ODR is used to estimate the γ coefficients, rather than OLS.

Model Outputs.

The model outputs include:

- Estimates of the actual amount of the contribution and the fractional contribution of the aerosol species of interest by the source or source type of interest to the receptor, along with the associated uncertainty estimates.
- Estimates of the average amount and the average fractional amount of the aerosol species of interest contributed by each source or source type of interest along with the associated uncertainty estimates.