Atmospheric Pollution Research

www.atmospolres.com

Source apportionment of $PM_{2.5}$ inside two diesel school buses using partial least squares discriminant analysis with chemical mass balance

Timothy Larson [1,2], Barbara Zielinska [3], Rob Ireson [4], L.J. Sally Liu [2,5,6]

[1] *Department of Civil and Environmental Engineering, Box 352700 University of Washington, Seattle, WA 98195*
[2] *Department of Environmental and Occupational Health Sciences, University of Washington*
[3] *Atmospheric Sciences Division, Desert Research Institute, Reno, NV.*
[4] *Air Quality Management Consulting, Greenbrae, CA.*
[5] *Swiss Tropical and Public Health Institute*
[6] *University of Basel, Basel, Switzerland*

ABSTRACT

Uncertainty–weighted partial least squares discriminant analysis was used to identify key species that were subsequently included in the EPA CMB8.2 chemical mass balance model to assess $PM_{2.5}$ source contributions from a previously published data set on school bus self–pollution. Estimates from this two–step modeling approach, herein referred to as effective variance discriminant analysis chemical mass balance (EVDA–CMB) were compared for eight separate runs with independent estimates from a synthetic tracer method. EVDA–CMB model predictions agreed favorably with those from the tracer method (R^2 = 0.83, 0.96 and 0.48, for contributions from the bus tailpipe, the engine crankcase and from other sources, respectively). Predictions from the traditional CMB model (without prior species selection), did not agree as well with the tracer method estimates of the bus tailpipe and engine crankcase contributions (R^2 = 0.18, 0.69, respectively), but did agree as well with the contributions from other sources (R^2 = 0.60). Although this study required discrimination of only a few sources, the same approach could be applied to the more general receptor modeling problem as an initial screening procedure, including approaches that optimize the choice of variables based on ambient data. This is important given that the number of species available for use in receptor modeling is rapidly expanding.

Keywords:
Chemical mass balance
Partial least squares discriminant analysis
Source apportionment

Article History:
Received: 18 June 2010
Revised: 23 December 2010
Accepted: 19 January 2011

Corresponding Author:
Timothy Larson
Tel: +1-206-543-6815
Fax: +1-206-685-3836
E-mail: tlarson@uw.edu

doi: 10.5094/APR.2011.019

1. Introduction

Receptor modeling has been used to estimate the contribution of various sources to measured airborne particulate matter concentrations (Henry, 1997; Hopke and Song, 1997; Seigneur et al., 1997). Traditionally, the U.S. EPA has recommended using the effective variance weighted chemical mass balance (CMB) receptor model (Miller et al., 1972; Watson et al., 1984), although less constrained multi–variate approaches have recently been widely used (Paatero, 1997; Henry et al., 1999; Paatero, 1999; Henry, 2003). More recent applications of the CMB method have explored the use of unique particulate organic tracers (Schauer et al., 1996; Zheng et al., 2002) as well as combined particulate and gaseous tracers (Schauer and Cass, 2000; Schauer et al., 2002).

One less well known alternative to CMB is partial least squares regression (PLS). PLS was originally developed by Herman Wold (Wold, 1966; Wold, 1981) and took his name when it was applied to the over–determined regression problem (Wold et al., 1983; Geladi and Kowalski, 1986). It was first applied to the aerosol source apportionment problem by Frank and Kowalski (1985). Vong et al. (1988) showed how PLS could solve this apportionment as a discriminant analysis problem. This latter approach has since been used in a limited number of similar studies (Larson and Vong, 1989; Vong, 1993; Wang and Larson, 1993; Norris, 1998). Similar to the effective variance–weighting scheme used in the EPA's CMB model (Watson et al., 1984), Norris (1998) introduced the idea of uncertainty weighted PLS, thereby accounting for individual species measurement uncertainties.

Here we apply uncertainty weighted PLS in order to determine key tracer species for subsequent use in a traditional chemical mass balance (CMB) model in order to estimate the source contributions to $PM_{2.5}$ inside a school bus. This two–step CMB model incorporating prior PLS discriminant analysis is one realization of what we refer to here as an effective variance discriminant analysis chemical mass balance model (EVDA–CMB).

The PLS algorithm provides an automated way to identify and highlight those species that differentiate the proposed sources, down–weighting the other species (Vong et al., 1988; Larson and Vong, 1989; Norris, 1998). These species are then used in CMB. The subsequent source contribution estimates are then compared with independent estimates of the relative contributions from each source that have been established by the use of unique, synthetic source tracers (Ireson et al., 2004; Zielinska et al., 2008; Liu et al., 2010).

Our data set is described in more detail elsewhere (Zielinska et al., 2008; Liu et al., 2010) and consists of ambient filter samples taken inside two diesel school buses and source samples taken from the tailpipe, from the crankcase road draft tube, and from the roadway traversed by each bus ("other sources"). Our initial attempts at CMB were only moderately successful in deducing the

relative source contributions to in–bus concentrations as judged by comparison with results obtained from the tracer–based method. This discrepancy was due in part to the relatively large number of measured species in this data set and the accompanying difficulty in selecting the appropriate species for use with CMB. We therefore decided to explore the use of an alternate species selection method for use with CMB, as described below.

2. Methods

The sampling and analysis methods are described in detail elsewhere (Zielinska et al., 2008; Liu et al., 2010) and briefly here. Unique, synthetic tracers were added to both the fuel supply and the lubricating oil. Tris(norbornadiene)iridium(III)acetylacetonate, an organometallic iridium complex was dissolved in toluene (1 g:225 mL) and added to each bus' fuel tank to track tailpipe exhaust particulate. Fully deuterated normal hexatriacontane ($n–C_{36}D_{74}$ or d–alkane) was dissolved in the bus' lubricant oil (100 g:18.9 L) to track crankcase emissions.

Source sampling involved using an on–board dilution tunnel (Weaver and Petty, 2004) to collect $PM_{2.5}$ samples from the tailpipe and the crankcase, respectively, of each bus. A lead vehicle drove the same route as the bus, ahead of the bus by approximately 5 minutes. A set of source profiles to represent other sources was developed based upon computed mass fractions of the species measured on the lead vehicle samples. In addition, a total of eight in–bus/lead vehicle sample pairs (Teflon and quartz filters) were taken using identical UMd impactors at 120 L/min. The windows in the lead vehicle were wide open during all sampling runs. The concentrations of particulate organic compounds are described in detail by Zielinska and co–workers (Zielinska et al., 2008).

The uncertainties for the in–bus samples were taken directly from the reported analytical uncertainties. The measurement uncertainties for the XRF and OC/EC fractions were reported using standard EPA protocols. The analytical uncertainties for the organic species were based on known deuterated internal standards. Compounds for which authentic standards were not available were quantified based on the response factor of standards most closely matched in structure and retention characteristics (Zelinksa et al., 2008). There were three sets of source samples taken for each of the eight runs (with one sample excluded due to sampling issues). The average analytical uncertainties of the three samples taken during each run were used as the uncertainties in this analysis.

2.1. CMB diagnostics

The standard EPA model, CMB8.2, was used in this analysis. It employs a weighted ordinary least squares solution to the following mass balance equation

$$C = FS + \varepsilon \tag{1}$$

where C (nx1) is the vector of observed concentrations of n species (μg/m^3), F (nxp) is a source profile matrix of n species from p sources (μg/μg mass), S (px1) is the source contribution vector (μg mass/m^3), and ε (nx1) is the vector of random measurement errors. The species are weighted by their respective measurement uncertainties involving an iterative procedure that includes the one standard deviation measurement uncertainties for the i^{th} species in both the source and ambient samples, σ_{source} (μg$_i$ /μg mass) and σ_{amb} (μg mass/m^3) respectively (Watson et al., 1984). Specifically, the weighted equation that is actually solved is:

$$C_W = F_W S + \varepsilon' \tag{2}$$

where

$$C_W = \left(V_e\right)^{-0.5} C \tag{3}$$

and

$$F_W = \left(V_e\right)^{-0.5} F \tag{4}$$

V_e (nxn) is the diagonal effective variance matrix whose off–diagonal elements are zero and whose diagonal elements are:

$$v_{e_{ii}} = \left(\sigma_{amb}\right)_i^2 + \sum_{j=1}^{p} \left(\sigma_{source}\right)_{i,j}^2 \left(s_j\right)^2 \tag{5}$$

where s_j is the contribution from the j^{th} source. The fact that the s_j are computed from Equation (2) means that the CMB8.2 algorithm is implicit and thus iterative. The first iteration initially assumes all the s_j are zero in Equation (5) and then computes the s_j from Equations (2)–(4) for use in subsequent iterations (Watson et al., 1984). The iteration procedure is stopped when the current and prior value of s_j are within one percent of each other. The final source contribution estimates in the original mass concentration units are then computed as:

$$S = F_W^+ C_W = \left(F^t \left(V_e\right)^{-1} F\right)^{-1} F^t \left(V_e\right)^{-1} C \tag{6}$$

By definition, the modified pseudo–inverse matrix (MPIN) is given as:

$$MPIN = F_W^+ = \left(F^t \left(V_e\right)^{-1} F\right)^{-1} F^t \left(V_e\right)^{-0.5} \tag{7}$$

Guidance is provided within CMB8.2 on those species that are influential and thus should be included in the model. Specifically, the elements of the normalized MPIN matrix, whose values range from –1 to 1, should be greater than 0.5 for species that are to be retained in the model (Kim and Henry, 1999; Watson, 2004).

Additional run diagnostics in CMB8.2 provide measures of the collinearity of the given set of weighted source profiles, including Henry's (1992) eligible space based on the singular value decomposition of the weighted F matrix as follows:

$$\left(V_e\right)^{-0.5} F = ADV^t \tag{8}$$

where A (n x n) and V (p x p) are orthogonal matrices and D is a diagonal matrix with p nonzero and positive elements called the singular values of the decomposition. V is the matrix of eigen–vectors of the decomposition. The eligible space is that spanned by these eigenvectors with inverse singular values less than or equal to the maximum score uncertainty. The estimable sources are those with a user defined minimum source projection within the estimable space, set at a default value of 0.95. CMB8.2 provides suggestions for combining highly collinear profiles (Henry, 1992), but provides no additional guidance on species selection so as to minimize collinearity of existing source profiles. Several authors have suggested alternative methods to minimize the collinearity problem, including ridge regression (Hopke, 1985) and non–negative principal component regression (Shi et al., 2009).

2.2. Species selection based on effective variance weighted discriminant analysis

As an alternate species selection strategy, we present here an effective variance weighted, partial least–squares discriminant analysis algorithm to select influential species for inclusion in the CMB model (EVDA–CMB) while minimizing collinearity. Source contributions from the crankcase, the tailpipe and other sources

(as captured by the lead vehicle measurements) predicted from the standard effective variance weighted CMB model (EPA–CMB8.2) and the alternate EVDA–CMB model were then compared with each other and with the source contribution estimates using the dual tracer method.

The PLS discriminant analysis model requires the identical information used in traditional CMB, namely a source profile matrix, F^t (pxn), and an ambient measurement vector, C (1xn), along with their associated measurement uncertainties. It also requires an identity matrix, Y (p x p). Both F^t and Y are decomposed into independent factors in a bilinear model that includes loadings, T, and scores, P and Q, as follows:

$$F^t = TP^t + E \tag{9}$$

$$Y = UQ^t + F^* \tag{10}$$

E and F^* are the residuals associated with the model fit. The solution is constrained in that T is orthogonal to P^t and U is orthogonal to Q^t. The solution provides a set of regression coefficients relating each factor of U uniquely with each factor of T (9 relationships in our case of a three source model). Instead of trying to maximize the variance, V, within F^t, as is done in principal component analysis (Shi et al., 2009), or maximizing the correlation, R^2, between F^t and Y as is done in multiple linear regression (MLR), PLS seeks to maximize the product V^*R^2 (Davies and Fearn, 2005). The PLS solution gives a different set of values for T and P than those derived from PCA or MLR (Barker and Ravens, 2003).

We use the multi–block PLS software provided for free by KVL (http://www.models.kvl.dk/source/). The species mass fractions from all source measurements were averaged into a single set of source profiles, F^t. Once the PLS algorithm finds a solution to Equations (9) and (10), if the diagonal elements of Y are near unity and the off–diagonal elements are near zero (minimal collinearity in the source profiles; model R^2 near one), then the model can be used to select species for inclusion into CMB.

The procedure described above does not account for different species measurement uncertainties. To do this, we need to include not only the uncertainties in the source profiles contained in F^t, but also the uncertainties in the ambient measurements contained in C. We do this with an effective variance weighting algorithm modeled after that used in EPA's CMB model (Watson et al., 1984). This recursive algorithm is described by Norris (1998) and shown below (the superscript k denotes the k^{th} iteration sequence):

Step 1: Convert the ambient measurements and their uncertain–ties into mass fractions

$$C_i' = \frac{C_i}{C_i^{mass}} \tag{11}$$

and

$$\left(\sigma_{amb}'\right)_i = \frac{\left(\sigma_{amb}\right)_i}{C_i^{mass}} \tag{12}$$

Step 2: Initially set $\left(\sigma_{eff}'\right)_i = \left(\sigma_{amb}'\right)_i$

Step 3: Weight the original source profiles by $\left(\sigma_{eff}'\right)_i$

$$\left(f_{i,j}^t\right)' = \frac{f_{i,j}^t}{\left(\sigma_{eff}'\right)_i} \tag{13}$$

where the $f_{i,j}$ are the elements of F and $\left(f_{i,j}^t\right)'$ are the elements of $(F^t)'$.

Step 4: Solve Equations (9) and (10) via PLS with $F^t = (F^t)'$

Step5: Using the PLS solution, i.e., the internal relationships between U and T, predict y by substituting C' for $(F^t)'$, where y is a p x 1 vector of predicted fractional contributions from each source to the ambient sample

Step 6: If $y_j < 0$, then $y_j = 0$ where y_j is the jth element of y.

Step 7: If k=1, go to *Step* 8;

Else if $\left|\hat{y}_j^k - \hat{y}_j^{k-1}\right| < 0.01$ for all j, then STOP

Else go to *Step* 8.

Step 8: Compute $\hat{y}_{norm}$ by scaling the p elements of the $\hat{y}$ vector such that $\sum_{j=1}^{p}\left(y_j\right)_{norm} = 1$

Step 9: Compute the effective measurement uncertainty similar to Equation (5) but now on a mass fraction basis as:

$$\sigma_{eff_i}' = \sqrt{\left(\sigma_{amb}'\right)_i^2 + \sum_{j=1}^{p}\left(\sigma_{source}\right)_{i,j}^2\left(\left(y_j\right)_{norm}\right)^2} \tag{14}$$

Step 10: Go to *Step* 3 and repeat the iteration sequence

To estimate the relative importance of each species in distinguishing a given source, we examined the elements of B_{pls} (nxp) that relate the scaled source profiles, $(F^t)'$ (pxn) to the discriminant matrix Y (pxp), where

$$\hat{Y} = \left(F^t\right)' B_{pls} \tag{15}$$

B_{pls} (nxp) can be computed from the PLS solution (Chong and Jun , 2005) as follows:

$$B_{pls} = W(P^tW)^{-1}(T^tT)^{-1}T^tY \tag{16}$$

where T (pxp), P (nxp) and W (nxp) are provided by the PLS algorithm such that $T = X(P^tW)^{-1}$ and Y (pxp) is the original identity matrix.

The major species identified by this discriminant model were then included in the CMB8.2 model by supplying truncated source profiles and ambient in–bus samples, considerably reducing the number of candidate species used while at the same time enhancing their discriminating power as source tracers. There are no currently universally accepted criteria for setting B_{pls} cutoff values (Chong and Jun, 2005; Anzanello et al., 2009). The species selection criteria used here is informal, choosing q species for the j^{th} source based on $(b_{pls})_{i,j}$, the individual elements of B_{pls}. For the j^{th} source (j^{th} column of B_{pls}), we chose those q species with:

$$\left|\left(b_{pls}\right)_{i,j}\right| > 0.5 * \max\left\{\left|\left(b_{pls}\right)_{i,j}\right|\right\} \text{ if } q>3 \quad (17)$$

otherwise we chose those *q* species with

$$\left|\left(b_{pls}\right)_{i,j}\right| > 0.1 * \max\left\{\left|\left(b_{pls}\right)_{i,j}\right|\right\} \text{ if } q\leq 3 \quad (18)$$

3. Results

For the EV–CMB runs, we used all 101 measured species. Detailed emissions rates and concentrations are reported previously (Zielinska et al., 2008; Liu et al., 2010). For all the partial least squares discriminant analysis predictions used to select species for the EVDA–CMB runs, the off–diagonal elements of *Y* predicted by PLS were near zero and model R^2 values were near 1.0: (range 0.999 to 0.9999). Table 1 lists the species with relatively high b_{pls} values that were used in the subsequent EVDA–CMB runs. As shown, the PLS procedure selected about 20 species for each run, with some species in common across all runs and others unique to a subset of runs.

Table 2 compares the estimated $PM_{2.5}$ source contributions using both EV–CMB and EVDA–CMB. In two cases the EV–CMB model failed to converge after 20 iterations. Table 3 summarizes selected CMB diagnostics for both models that are relevant to source profile collinearity (the number of estimable sources and the maximum inverse singular value). Also shown are the number of important fitting species as identified by the MPIN matrix with elements >0.5. In all cases where EV–CMB ran successfully, there was no obvious indication of collinearity issues.

Figure 1 compares the initial CMB model predictions with the dual tracer (DT) method estimates of the crankcase, tailpipe and "other" contributions. These latter DT estimates were previously reported (Liu et al., 2010). As shown, the EVDA–CMB model shows much better agreement with the DT method than the traditional EV–CMB model.

Table 1. *Value of $(b_{pls})_{i,j}$ for each selected species and for each sample as determined from Equations (16)–(18)*

Species[a]	Source[b]	Sample[c]							
		B1C AM	B1C PM	B1O AM	B1O PM	B2C AM	B2C PM	B2O AM	B2O PM[d]
Hexadecylcyclohexane	CK			0.010					0.027
Dotriacontane	CK					0.007			
Hopane 13	CK	0.018	0.015	0.011	0.013	0.019	0.028	0.024	0.033
Hopane 17	CK	0.018	0.016	0.020	0.018	0.019	0.029	0.031	0.058
Hopane 19	CK	0.018	0.015	0.013	0.014	0.017	0.024	0.024	0.035
Hopane 21	CK	0.018	0.016	0.015	0.015	0.019	0.027	0.026	0.042
Hopane 22	CK	0.018	0.015	0.013	0.015	0.019	0.026	0.027	0.038
Hopane 24	CK		0.014	0.010	0.011	0.017	0.022	0.020	0.028
Hopane 25	CK	0.016	0.014	0.008	0.011	0.017	0.025	0.020	0.021
Hopane 26	CK					0.017			
OC1	CK	0.016		0.012		0.017	0.024	0.023	0.038
OC2	CK	0.017	0.015	0.011		0.018	0.023	0.022	0.030
OC	CK	0.016	0.014			0.018	0.023	0.021	0.028
Sterane 43	CK					0.016			
Sterane 44	CK	0.016	0.015	0.009	0.011	0.017	0.024	0.020	0.025
Sterane 45	CK	0.017	0.015	0.009	0.011	0.018	0.024	0.021	0.027
Sterane 47	CK						0.021		
Sterane 48	CK		0.015						
Sterane 50	CK							0.019	
Sterane 51	CK		0.015	0.011	0.013		0.023	0.024	0.031
Sterane 52	CK	0.017			0.011				
Triacontane	CK								0.020
EC1	TP	0.012	0.012		0.020	0.008	0.026	0.030	0.046
EC2	TP	0.021	0.019	0.023	0.029	0.024	0.036	0.043	0.079
EC	TP	0.020	0.019	0.026	0.033	0.018	0.037	0.047	0.081
Eicosane	TP					0.007			
Heptadecane	TP			0.017	0.031		0.022		0.039
Hexadecane	TP			0.008		0.004			
Hexatriacontane	TP	0.012	0.009						
Nonadecane	TP	0.010	0.009	0.022	0.031	0.010	0.029	0.035	0.061
Norpristane	TP				0.021				
Octadecane	TP	0.008	0.008	0.015	0.032	0.008	0.028		0.036
Phytane	TP				0.022				
Farnesane	LV			0.009	0.007		0.009		
Norfarnesane	LV							0.004	
Sulfur	LV	0.014	0.014	0.039	0.035	0.011	0.020	0.033	0.052
Tricosane	LV			0.007	0.006				0.005

[a] Hopane13: 18α(H)–22,29,30–Trisnorneohopane; Hopane17: 17α(H),21ß(H)–30–Norhopane; Hopane19: 17α(H),21ß(H)–Hopane; Hopane21: 22S–17α(H),21ß(H)–30–Homohopane; Hopane22: 22R–17α(H),21ß(H)–30–Homohopane; Hopane24: 22S–17α(H),21ß(H)–30,31–Bishomohopane; Hopane25: 22R–17α(H),21ß(H)–30,31–Bishomohopane; Hopane26: 22S–17α(H),21ß(H)–30,31,32–Trisomohopane; Sterane43: 20R–5α(H),14ß(H),17ß(H)–cholestane; Sterane44: 20S–5α(H),14ß(H),17ß(H)–cholestane; Sterane45: 20R–5α(H),14α(H),17α(H)–cholestane & 20S–13ß(H),17α(H)–diastigmastane; Sterane47: 20R–5α(H),14ß(H),17ß(H)–ergostane; Sterane48: 20S–5α(H),14ß(H),17ß(H)–ergostane & 20R–13α(H),17ß(H)–diastigmastane; Sterane 50: 20S–5α(H),14α(H),17α(H)–stigmastane; Sterane51: 20R–5α(H),14ß(H),17ß(H)–stigmastane; Sterane52: 20S–5α(H),14ß(H),17ß(H)–stigmastane.

[b] CK = crankcase; TP = tailpipe; LV = lead vehicle

[c] B1: bus 1; B2: bus 2; C: windows closed; O: windows open; AM: morning sampling; PM: afternoon sampling

[d] the LV profile from the corresponding AM run was used due to problems with the filter mass value for the PM run

Table 2. *Comparison of EVDA–CMB and EV–CMB predictions of the source contribution estimates inside the bus for emissions from the bus tailpipe (TP), bus crankcase (CK) and from other sources as indicated by the lead vehicle (LV) measurements (units are μg/m³ $PM_{2.5}$)*

Sample[a]	EVDA–CMB			EV–CMB		
	TP	CK	LV	TP	CK	LV
B1C AM	1.4 *(1.1)*[b]	12.3 *(2.8)*	5.6 *(0.3)*	–[c]	–	–
B1C PM	0 *(0.6)*	16.7 *(3.3)*	26.4 *(2.1)*	0 *(0.02)*	0 *(0.01)*	31.0 *(1.0)*
B1O AM	0.9 *(0.6)*	1.8 *(0.4)*	6.9 *(0.1)*	6.6 *(1.2)*	1.4 *(0.3)*	6.8 *(0.1)*
B1O PM	0.7 *(0.4)*	1.3 *(0.3)*	7.2 *(0.2)*	5.9 *(0.9)*	1.6 *(0.3)*	7.0 *(0.2)*
B2C AM	3.8 *(0.6)*	27 *(3.4)*	3.7 *(0.4)*	0 *(0.1)*	0 *(0.1)*	7.5 *(0.2)*
B2C PM	3.2 *(0.5)*	16.4 *(2.2)*	5.2 *(0.5)*	–	–	–
B2O AM	1.0 *(0.2)*	3.7 *(0.6)*	7.1 *(0.5)*	0 *(0.1)*	0 *(0.1)*	12.2 *(0.4)*
B2O PM	1.5 *(0.3)*	5.0 *(0.8)*	13.4 *(0.8)*	0 *(0.1)*	0 *(0.1)*	17.8 *(0.5)*

[a] *B1: bus 1; B2: bus 2; C: windows closed; O: windows open; AM: morning sampling; PM: afternoon sampling*
[b] *( ) = standard error as estimated by CMB*
[c] *solution did not converge in 20 iterations*

Figure 1 compares the initial CMB model predictions with the dual tracer (DT) method estimates of the crankcase, tailpipe and "other" contributions. These latter DT estimates were previously reported (Liu et al., 2010). As shown, the EVDA–CMB model shows much better agreement with the DT method than the traditional EV–CMB model.

4. Discussion

This study offered the unique opportunity to compare the predictions from two different versions of the CMB model to those using unique, synthetic tracers. The lead–vehicle approach minimized the total number of source profiles, optimizing the receptor model's chances of success. Although the EVDA–CMB model performed better than the traditional EV–CMB model, it is possible that further species selection efforts could have improved the latter model. However, there were no obvious indications from the EV–CMB diagnostics of how to best proceed with species selection. The PLS predictions of the discriminant matrix, *Y*, provides additional important information on the potential collinearities of chosen source profiles. In our example, these profiles proved to be adequate separated, in agreement with the standard CMB diagnostics. The $\hat{y}$ vector also provides estimates of the fractional source contributions to the ambient sample independent of CMB8.2. However, given that the EVDA algorithm re–normalizes $\hat{y}$ at each iteration step to assist convergence, the final $\hat{y}$ values could be somewhat biased. This needs further exploration, perhaps with an appropriate artificial data set.

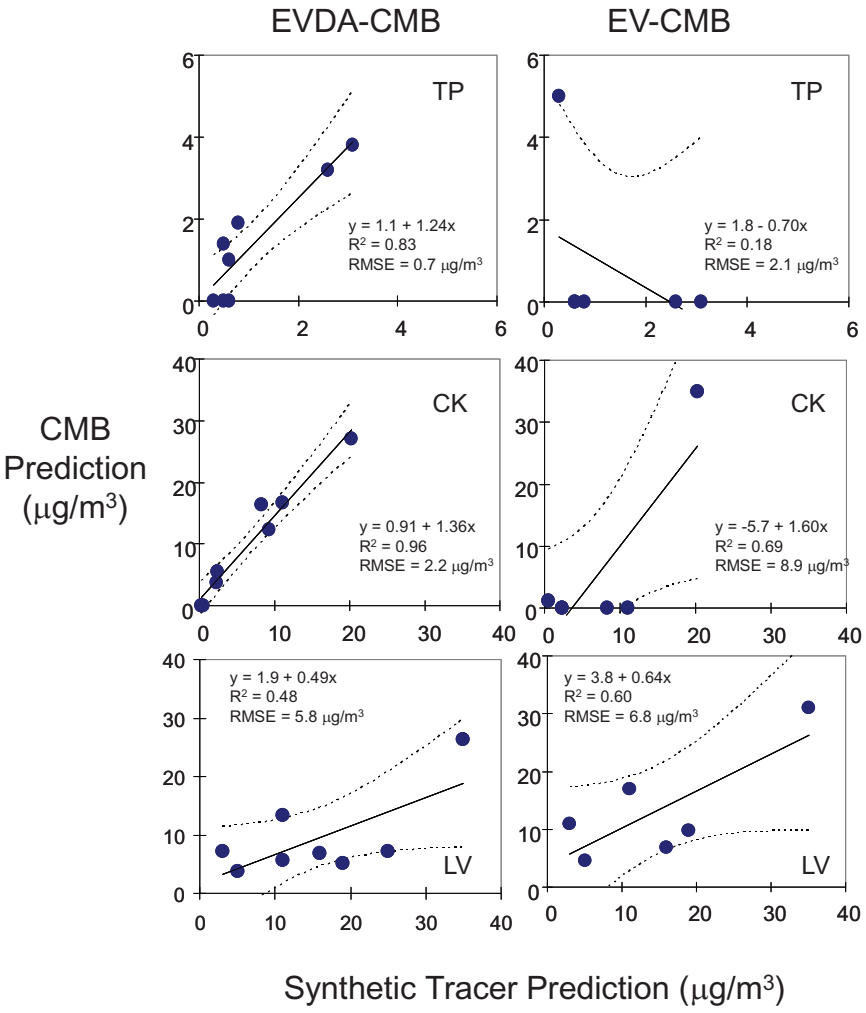

Figure 1. *Comparison of EVDA–CMB and EV–CMB predictions of $PM_{2.5}$ source contributions with those from the dual tracer method (TP = tailpipe, CK = crankcase, LV = lead vehicle).*

Table 3. *Selected CMB diagnostics from model runs using all measured species (EV–CMB) and species selected by partial least squares discriminant analysis (EVDA–CMB)*

CMB Diagnostics		Sample [a]							
		B1C AM	B1C PM	B1O AM	B1O PM	B2C AM	B2C PM	B2O AM	B2O PM[b]
EV–CMB (all species)									
Number of Estimable Sources		0	3	3	3	0	3	3	3
Maximum [(singular value)$^{-1}$]		–	1.0	1.2	0.9	–	0.3	0.4	0.6
Number of species with |MPIN| > 0.5	TP	–	0	2	3	–	0	0	0
	CK	–	0	2	2	–	0	0	0
	LV	–	5	1	1	–	1	2	2
EVDA –CMB (selected species[c])									
Number of Estimable Sources		3	3	3	3	3	3	3	3
Maximum [(singular value)$^{-1}$]		2.8	3.5	0.6	0.5	3.5	2.2	0.6	0.9
Number of species with |MPIN| > 0.5	TP	3	2	6	5	3	3	5	4
	CK	13	18	14	11	18	15	15	14
	LV	1	1	1	1	1	1	1	1

[a] *B1: bus 1; B2: bus 2; C: windows closed; O: windows open; AM: morning sampling; PM: afternoon sampling*
[b] *( ) = standard error as estimated by CMB*
[c] *solution did not converge in 20 iterations*

Although this study required discrimination of only a few sources, the same approach could be applied to the more general receptor modeling problem as an initial screening procedure, including approaches that optimize the choice of variables based on ambient data (Marmur et al., 2007). This is important given that the number of species available for use in receptor modeling is rapidly expanding with the continuous improvements in analytical organic chemistry.

Acknowledgements

This study was partially sponsored by the National Institute of Environmental Health Sciences (#1R01ES12657–01A1), a gift fund from the International Truck and Engine Corporation with the University of Washington, and the U.S. Department of Energy Office of FreedomCAR and Vehicle Technologies through the National Renewable Energy Laboratory. We thank the technical supports from David Anderson of the Seattle School District Transportation Department, First Student, Inc., and the Puget Sound Clean Air Agency.

References

Anzanello, M.J., Albin, S.L., Chaovalitwongse, W.A., 2009. Selecting the best variables for classifying production batches into two quality levels. *Chemometrics and Intelligent Laboratory Systems* 97, 111–117.

Barker, M., Ravens, W., 2003. Partial least squares for discrimination. *Journal of Chemometrics* 17, 166–173.

Chong, I.G., Jun, C.H., 2005. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* 78, 103–112.

Davies, A.M.C., Fearn, T., 2005. Back to basics: observing PLS. *Spectroscopy Europe* 17, 28–29.

Frank, I.E., Kowalski, B.R., 1985. Statistical Receptor Models Solved by Partial Least Squares. *Environmental Applications of Chemometrics* 292, 271–279.

Geladi, P., Kowalski, B.R., 1986. Partial least squares regression: a tutorial. *Analytica Chimica Acta* 185, 1–17.

Henry, R.C., 2003. Multivariate receptor modeling by N–dimensional edge detection. *Chemometrics and Intelligent Laboratory Systems* 65, 179–189.

Henry, R.C., Park, E.S., Spiegelman, C.H., 1999. Comparing a new algorithm with the classic methods for estimating the number of factors. *Chemometrics and Intelligent Laboratory Systems* 48, 91–97.

Henry, R.C., 1997. History and fundamentals of multivariate air quality receptor models. *Chemometrics and Intelligent Laboratory Systems* 37, 37–42.

Henry, R.C., 1992. Dealing with near collinearity in chemical mass balance receptor models. *Atmospheric Environment* 26A, 933–938.

Hopke, P.K., Song, X., 1997. The chemical mass balance as a multivariate calibration problem. *Chemometrics and Intelligent Laboratory Systems* 37, 5-14.

Hopke, P.K., 1985. *Receptor Modeling in Environmental Chemistry*. John Wiley and Sons, Inc., New York, pp. 132–140.

Ireson, R.G., Easter, M.D., Lakin, M.L., Ondov, J.M., Clark, N.N., Wright, D.B., 2004. Estimation of diesel particulate matter concentrations in a school bus using a fuel–based tracer: a sensitive and specific method for quantifying vehicle contributions. *Transportation Research Record* 1880, 21–28.

Kim, B.M., Henry, R.C., 1999. Diagnostics for determining influential species in the chemical mass balance receptor model. *Journal of the Air & Waste Management Association* 49, 1449-1455.

Larson, T.V., Vong, R.J., 1989. Partial least squares regression methodology: application to source receptor modeling. *Receptor Models in Air Resources Management, Air and Waste Management Association*, pp. 391- 397.

Liu, L.J.S., Phuleria, H.C., Webber, W., Davey, M., Lawson, D.R., Ireson, R.G., Zielinska, B., Ondov, J.M., Weaver, C.S., Lapin, C.A., Easter, M., Hesterberg, T.W., Larson, T., 2010. Quantification of self–pollution from two diesel school buses using three independent methods. *Atmospheric Environment* 44, 3422–3431.

Marmur, A., Mulholland, J.A., Russell, A.G., 2007. Optimized variable source–profile approach for source apportionment. *Atmospheric Environment* 41, 493–505.

Miller, M.S., Friedlander, S.K., Hidy, G.M., 1972. A chemical element balance for the Pasadena aerosol. *Journal of Colloidal and Interface Science* 39, 165–176.

Norris, G., 1998. *Air Pollution and the Exacerbation of Asthma in an Arid, Western U.S. City*. Ph.D. Thesis, University of Washington, Seattle, United States, 199 pp.

Paatero, P., 1999. The multilinear engine – a table–driven, least squares program for solving multilinear problems, including the n–way parallel factor analysis model. *Journal of Computational and Graphical Statistics* 8, 854–888.

Paatero, P., 1997. Least squares formulation of robust non–negative factor analysis. *Chemometrics and Intelligent Laboratory Systems* 37, 23–35.

Schauer, J.J., Fraser, M.P., Cass, G.R., Simoneit, B.R.T., 2002. Source reconciliation of atmospheric gas–phase and particle–phase pollutants during a severe photochemical smog episode. *Environmental Science and Technology* 36, 3806–3814.

Schauer, J.J., Cass, G.R., 2000. Source apportionment of wintertime gas–phase and particle–phase air pollutants using organic compounds as tracers. *Environmental Science and Technology* 34, 1821–1832.

Schauer, J.J., Rogge, W.F., Hildemann, L.M., Mazurek, M.A., Cass, G.R., Simoneit, B.R.T., 1996. Source apportionment of airborne particulate matter using organic compounds as tracers. *Atmospheric Environment* 30, 3837–3855.

Seigneur, C., Pai, P., Louis, J.F., Hopke, P., Grosjean, D., 1998. Review of Air Quality Models for Particulate Matter. American Petroleum Institute Document No. CP015–97–1b.

Shi, G.L., Feng, Y.C., Zeng, F., Li, X., Zhang, Y.F., Wang, Y.Q., Zhu, T., 2009. Use of a nonnegative constrained principal component regression chemical mass balance model to study the contributions of nearly collinear sources. *Environmental Science and Technology* 43, 8867–8873.

Vong, R.J., 1993. Atmospheric chemometrics for identification of trace element sources in precipitation. *Analytica Chimica Acta* 277, 389–404.

Vong, R., Geladi, P., Wold, S., Esbensen, K., 1988. Source contributions to ambient aerosol calculated by discriminant partial least squares regression (PLS). *Journal of Chemometrics* 2, 281–296.

Wang, S.Z., Larson, T., 1993. Ambient error weighted partial least squares regression: a new receptor model. *Analytica Chimica Acta* 272, 333–337.

Watson, J.G., 2004. Protocol for Applying and Validating the CMB Model for $PM_{2.5}$ and VOC. U.S.EPA Report EPA–451/R–04–001.

Watson, J.G., Cooper, J.A., Huntzicker, J.J., 1984. The effective variance weighting for least squares calculations applied to the mass balance receptor model. *Atmospheric Environment* 18, 1347–1355.

Weaver, C.S., Petty, L.E., 2004. Reproducibility and Accuracy of On–Board Emissions Measurements Using the RAVEM System. *SAE International*, SAE Paper No. 2004–01–0965.

Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration problem in chemistry solved by the PLS method. *Matrix Pencils Lecture Notes in Mathematics*, Springer, Heidelberg, pp.286-293.

Wold, H., 1981. *Systems under Indirect Observation, Causality–Structure–Prediciton*, Amsterdam, pp. 1–54.

Wold, H., 1966. *Multivariate Analysis*. Academic Press**,** New York, pp. 391–420.

Zielinska, B., Campbell, D., Lawson, D.R., Ireson, R.G., Weaver, C.S., Hesterberg, T.W., Larson, T., Davey, M., Liu, L.J.S., 2008. Detailed characterization and profiles of crankcase and diesel particulate matter exhaust emissions using speciated organics. *Environmental Science and Technology* 42, 5661–5666.

Zheng, M., Cass, G.R., Schauer, J.J., Edgerton, E.S., 2002. Source apportionment of $PM_{2.5}$ in the southeastern United States using solvent–extractable organic compounds as tracers. *Environmental Science and Technology* 36, 2361–2371.