APPENDIX 6D: Chemical Mass Balance Receptor Model

Introduction

Receptor models use the chemical and physical characteristics of gases and particles measured at source and receptor to both identify the presence of and to quantify source contributions to the receptor. The particle characteristics must be such that: 1) they are present in different proportions in different source emissions; 2) these proportions remain relatively constant for each source type; and 3) changes in these proportions between source and receptor are negligible or can be approximated.

Common types of receptor models include: 1) chemical mass balance (CMB); 2) principal component analysis (PCA), otherwise known as factor analysis); and 3) multiple linear regression (MLR). Extensive explanations of each of these models, operating separately and together, are given by Watson,[1] Chow,[2] Hopke,[3] and Watson.[4, 5] The PCA, CMB, and MLR have been combined with a dispersion model in a PM_{10} assessment package prepared for the California Air Resources Board[4, 6] which provides interfaces among data bases and modeling software. The chemical mass balance (CMB) is the fundamental receptor model, and the derivation of the PCA and MLR models from fundamental physical principles begins with the CMB.

The chemical mass balance consists of a least squares solution to a set of linear equations which expresses each receptor concentration of a chemical species as a linear sum of products of source profile species and source contributions. The source profile species (i.e., the fractional amount of the species in the emissions from each source type) and the receptor concentrations, with appropriate uncertainty estimates, serve as input data to the CMB model. The output consists of the amount contributed by each source type to each chemical species. The model calculates values for the contributions from each source and the uncertainties of those values. Input data uncertainties are used both to weight the importance of input data values in the solution and to calculate the uncertainties of the source contributions.

Model Equations

Suppose our list of aerosol components includes only material that is nonreactive and maintains relative ratios between various species as material is transported from source to receptor. In this case $K_c(i,j,k)$ are all zero and $K_d(i,j,k)$ are the same for all elements i. Their common value is denoted by $K_d(j,k)$ indicating the nondependence on i. This implies that the quantities r_{ijk} do

not depend on i. Then,

$$\frac{C_{ijk}}{C_{i'jk}} = \frac{c_{ijk}r_{ijk}d_{jk}}{c_{i'jk}r_{i'jk}d_{jk}} = \frac{c_{ijk}}{c_{i'jk}} \quad (1)$$

which implies that the signature for source j at the source equals the signature for source j as perceived at the receptor.

Let $S_{jk} = \sum_{i=1}^{m} C_{ijk}$. The quantity S_{jk} is the concentration of the aerosol mixture at the receptor during sampling period k that is attributable to source j. The fraction a_{ijk} defined by

$$a_{ijk} = \frac{C_{ijk}}{S_{jk}} \quad (2)$$

is then the fraction of the aerosol mixture at the receptor attributable to source j that is species i, during sampling period k. Thus the numbers a_{ijk} for $i = 1, 2, \ldots, m$, represent the source signature for source j for sampling period k. If the a_{ijk} for all the sources affecting the receptor site are known then the following set of simultaneous equations holds.

$$C_{ik} = \sum_{i=1}^{n} a_{ijk}S_{jk} \quad (3)$$

in n unknowns $S_{1k}, S_{2k}, \ldots, S_{nk}$, for each of the sampling periods $k = 1, 2, \ldots, s$. Equation – is the chemical mass balance equation. The number of equations must be greater than or equal to the number of chemical species (I) for a unique solution to these equations.

Solutions to the CMB equations that have been used are:1) a tracer solution; 2) a linear programming solution; 3) an ordinary weighted least squares solution with or without an intercept; 4) a ridge regression weighted least squares solution with or without an intercept; and 5) an effective variance least squares solution with or without an intercept. An estimate of the uncertainty associated with the source contributions is an integral part of several of these solution methods.

Weighted linear least squares solutions are preferable to the tracer and linear programming solutions because: 1) theoretically they yield the most likely solution to the CMB equations providing model assumptions are met; 2) they can make use of all available chemical measurements, not just the so- called tracer species; 3) they are capable of analytically estimating the uncertainty of the source contributions;

CMB software in current use[11] applies the effective variance solution developed and tested by Watson[12] because this solution: 1) provides realistic estimates of the uncertainties of the source contributions (owing to its incorporation of both source profile and receptor data uncertainties); and 2) chemical species with higher precisions in both the source and receptor measurements are given greater influence than are species with lower precisions.

The effective variance solution is derived by minimizing the weighted sums of the squares of the differences between the measured and calculated values of C_i and F_{ij}.[13] The solution algorithm is an iterative procedure which calculates a new set of S_j based on the S_j estimated from the previous iteration.

Watson[7] found that individual sources with similar source profiles would yield unreliable values if included in the same chemical mass balance. Henry[15] proposed a quantitative method of identifying this interference between this similar source compositions, which is known as "collinearity." His "singular value decomposition" defines an "estimable space into which resolvable sources should lie." The sources which do not fall into this estimable space are collinear, or too similar to be resolved from the sources which do lie within the estimable space. Henry[15] further proposed that linear combinations of source contributions resulting from collinear source compositions would be more representative of the summed contributions of these sources.

Williamson and Dubose[14] claimed that the ridge regression reduces collinearities. Henry[15] tested the ridge regression solution with respect to the separation of urban and continental dust and found that the bias resulted in physically unrealistic negative values for several of the F_{ij}. The ridge regression solution has not been used in the CMB since these tests were published.

CMB Model Assumptions

The CMB model assumptions are:

1. Compositions of source emissions are constant over the period of ambient and source sampling.
2. Chemical species do not react with each other, i.e., they add linearly.
3. All sources with a potential for significantly contributing to the receptor have been identified and have had their emissions characterized.
4. The sources' compositions are linearly independent of each other.
5. The number of sources or source categories is less than or equal to the number of chemical species.
6. Measurement uncertainties are random, uncorrelated, and normally distributed.

Effects of Deviations from CMB Model Assumptions

Assumptions 1 through 6 for the CMB model are fairly restrictive and will never be totally complied with in actual practice. Fortunately, the CMB model can tolerate reasonable deviations from these assumptions, though these deviations increase the stated uncertainties of the source contribution estimates.

The CMB model has been subjected to a number of tests to determine its abilities to tolerate deviations from model assumptions[7, 16, 15, 17, 18, 19, 20, 21, 22, 23, 5] These studies all point to the same basic conclusions regarding deviations from the above-stated assumptions.

With regard to Assumption 1, source compositions, as seen at the receptor, are known to vary substantially among sources, and even within a single source over an extended period of time. These variations are both systematic and random and are caused by three phenomena: 1) transformation and deposition between the emissions point and the receptor; 2) differences in

fuel type and operating processes between similar sources or the same source in time; and 3) uncertainties or differences between the source profile measurement methods. Evaluation studies have generally compared CMB results from several tests using randomly perturbed input data and from substitutions of different source profiles for the same source type. The general conclusions drawn from these tests are:

- The error in the estimated source contributions due to biases in all of the elements of a source profile is in direct proportion to the magnitude of the biases.
- For random errors, the magnitude of the source contribution errors decreases as the number of components increases.

The most recent and systematic tests are those of Javitz[23] which apply to a simple four-source urban airshed and a complex ten-source urban airshed. These tests with 17 commonly measured chemical species showed that primary mobile, geological, coal-fired power plant, and vegetative burning source types can be apportioned with uncertainties of approximately 30% when coefficients of variation in the source profiles are as high as 50%. This performance was demonstrated even without the presence of unique "tracer" species such as selenium for coal-fired power plants or soluble potassium for vegetative burning. In a complex urban airshed, which added residual oil combustion, marine aerosol, steel production, lead smelting, municipal incineration, and a continental background aerosol, it was found that the geological, coal-fired power plant, and background source profiles were collinear with the measured species. At coefficients of variation in the source profiles as low as 25%, average absolute errors were on the order of 60%, 50%, and 130% for the geological, coal-burning, and background sources, respectively. All other sources were apportioned with average absolute errors of approximately 30% even when coefficients of variation in the source profiles reached 50%. Once again, these tests were performed with commonly measured chemical species, and results would improve with a greater number of species which are specifically emitted by the different source types.

With regard to the nonlinear summation of species, Assumption 2, no studies have been performed to evaluate deviations from this assumption. While these deviations are generally assumed to be small, conversion of gases to particles and reactions between particles are not inherently linear processes. This assumption is especially applicable to the end products of photochemical reactions and their apportionment to the sources of the precursors. Further model evaluation is necessary to determine the tolerance of the CMB model to deviations from this assumption. The current practice is to apportion the primary material which has not changed between source and receptor. The remaining quantities of reactive species such as ammonium, nitrate, sulfate, and elemental carbon are then apportioned to chemical compounds rather than directly to sources. While this approach is not as satisfying as a direct apportionment, it at least separates primary from secondary emitters and the types of compounds apportioned give some insight into the chemical pathways which formed them. As chemical reaction mechanisms and rates, deposition velocities, atmospheric equilibrium, and methods to estimate transport and aging time become better developed, it may be possible to produce "fractionated" source profiles which will allow this direct attribution or reactive species to sources. Such apportionment will require measurements of gaseous as well as particulate species at receptor sites.

A major challenge to the application of the CMB is the identification of the primary contributing sources for inclusion in the model, Assumption 3. Watson[7] systematically increased the number of sources contributing to his simulated data from four to eight contributors while solving the CMB equations assuming only four sources. He also included more sources in the least squares solutions than those which were actually contributors, with the following results:

- Underestimating the number of sources had little effect on the calculated source contributions if the prominent species contributed by the missing sources were excluded from the solution.
- When the number of sources was underestimated, and when prominent species of the omitted sources were included in the calculation of source contributions, the contributions of sources with properties in common with the omitted sources were overestimated.
- When source types actually present were excluded from the solution, ratios of calculated to measured concentrations were often outside of the 0.5 to 2.0 range, and the sum of the source contributions was much less than the total measured mass. The low calculated/measured ratios indicated which source compositions should be included.
- When the number of sources was overestimated, the sources not actually present yielded contributions less than their standard errors if their source profiles were significantly distinct from those of other sources. The over-specification of sources decreased the standard errors of the source contribution estimates.

Recent research suggests that Assumption 3 should be restated to specify that source contributions above detection limits should be included in the CMB. At this time, however, it is not yet possible to determine the "detection limit" of a source contribution at a receptor since this is a complicated and unknown function of the other source contributions, the source composition uncertainties and the uncertainties of the receptor measurements. Additional model testing is needed to define this "detection limit."

The linear independence of source compositions required by Assumption 4 has become a subject of considerable interest since the publication of Henry's[15] singular value decomposition (SVD) analysis. As previously noted, this analysis provides quantitative measures of collinearity and the sensitivity of CMB results to specific receptor concentrations. These measures can be calculated analytically in each application. Henry[15] also proposed an optimal linear combination of source contributions that have been determined to be collinear.

Other "regression diagnostics" have been summarized by Belsley[24] and have been applied to the CMB by DeCesar.[20, 21] Kim and Henry[25] show that most of these diagnostics are useless because they are based on the assumption of zero uncertainty in the source profiles. They demonstrate, through the examination of randomly perturbed model input data, that the values for these diagnostics vary substantially with typical random changes in the source profiles.

Tests performed on simulated data with obviously collinear source compositions typically result in positive and negative values for the collinear source types as well as large standard errors on the collinear source contribution estimates. Unless the source compositions are nearly identical, the sum of these large positive and negative values very closely approximates the sum of the true contributions.

With most commonly measured species (e.g., ions, elements and carbon) and source types (e.g., motor vehicle, geological, residual oil, sea salt, steel production, wood burning and various industrial processes), from five to seven sources are linearly independent of each other in most cases.[23]

Gordon[16] found instabilities in the ordinary weighted least square solutions to the CMB equations when species presumed to be "unique" to a certain source type were removed from the solution. Using simulated data with known perturbations ranging from 0 to 20 percent, Watson[7] found: "In the presence of likely uncertainties, sources such as urban dust and continental background dust cannot be adequately resolved by least squares fitting, even though their compositions are not identical. Several nearly unique ratios must exist for good separation."

With regard to Assumption 5, the true number of individual sources contributing to receptor concentrations is generally much larger than the number of species that can be measured. It is therefore necessary to group sources into source types of similar compositions so that this assumption is met. For the most commonly measured species, meeting Assumption 4 practically defines these groupings.

With respect to Assumption 6 (the randomness, normality, and the uncorrelated nature of measurement uncertainties), there are no results available from verification or evaluation studies. Every least squares solution to the CMB equations requires this assumption, as demonstrated by the derivation of Watson.[12] In reality, very little is known about the distribution of errors for the source compositions and the ambient concentrations. If anything, the distribution probably follows a log-normal rather than a normal distribution. Ambient concentrations can never be negative, and a normal distribution allows for a substantial proportion of negative values, while a log-normal distribution allows no negative values. For small errors (e.g., less than 20%), the actual distribution may not be important, but for large errors, it probably is important. A symmetric distribution becomes less probable as the coefficient of variation of the measurement increases. This is one of the most important assumptions of the solution method that requires testing.

Model Input and Output Data

The chemical mass balance modeling procedure requires: 1) identification of the contributing sources types; 2) selection of chemical species to be included; 3) estimation of the fraction of each of the chemical species which is contained in each source type (i.e., the source compositions); 4) estimation of the uncertainty in both ambient concentrations and source compositions; and 5) solution of the chemical mass balance equations, and 6) validation and reconciliation. Each of these steps requires different types of data.

Emissions inventories are examined to determine the types of sources which are most likely to influence a receptor. Principal components analysis applied to a time series of chemical measurements is also a useful method of determining the number and types of sources. After these sources have been identified, profiles acquired from similar sources[26] (identify most of the available source profiles) are examined to select the chemical species to be measured. Watson[7] demonstrates that the more species measured, the better the precision of the CMB apportionment.

The ambient concentrations of these species, C_i, and their fractional amount in each source-type emission, F_{ij}, are the measured quantities which serve as CMB model input data. These values

require uncertainty estimates, σ_{Ci} and σ_{Fij}, which are also input data. Input data uncertainties are used both to weight the importance of input data values in the solution and to calculate the uncertainties of the source contributions. The output consists of: 1) the source contribution estimates (S_j) of each source type; 2) the standard errors of these source contribution estimates. 3) the amount contributed by each source type to each chemical species.

Model Applications

Most of the applications of the CMB model have been directed at the apportionment of PM_{10} (mass of suspended particles in sizes less than 10 μm in aerodynamic diameter) and $PM_{2.5}$ (mass of suspended particles in sizes less than 2.5 μm in aerodynamic diameter) measurements and source apportionments are reported for cities in potential non-attainment areas of Alaska,[27] Arizona,[28] California,[29, 30, 27] Colorado,[31] Illinois,[32] and Nevada.[33, 34] Chow and Watson[26] summarize a large number of source and receptor data bases which have been compiled for CMB modeling.

The results of these studies show that receptor models provide results which allow regulatory decisions to be made about pollution control measures. With the appropriate chemical measurements, receptor models can apportion general source types such as primary geological, primary motor vehicle exhaust, primary vegetative burning, primary marine, and primary industrial (e.g., copper smelting, gypsum mining, steel production). These primary contributions can be separated from secondary sulfate and nitrate contributions. The apportionment of the secondary aerosol to emitters of its precursors was not attempted in any of these studies, though Lurmann[35] provides a theoretical framework for estimating the changes in the sulfate and nitrate proportions in source profiles which take place between source and receptor.

Ryan[28] and Dresser and Baird[31] show that receptor and dispersion model source contribution estimates can be reconciled when EPA protocols are followed. In each of these cases, the first application of source and receptor models returned widely varying results. When emissions source-types in both models were re-defined so that they represented the same combination of emitters, when emission rates were re-evaluated, and when new source profiles were obtained, the source apportionments from both receptor and dispersion models came into agreement within uncertainty intervals.

Magliano,[30] Hopke,[32] and Cheng,[36] applied factor analysis or principal component analysis (PCA) to aid in the identification of the contributing sources in the airsheds they studied. These authors show the value of these multivariate methods in identifying sources and estimating the chemical profiles of their primary particle emissions when a sufficient number of chemically characterized samples (greater than 60) are available.

The extension of CMB source apportionment of $PM_{2.5}$ to the apportionment of light extinction is fairly simple. The chemical components of $PM_{2.5}$ are the major cause of non–Rayleigh light extinction in both urban and nonurban areas. The relative importance of chemical species to light extinction is 1 for soil-related species, 3 to 4 for organic carbon, 5 to 6 for nitrates, 6 to 9 for sulfates, and 10 for elemental carbon. Nitrogen dioxide gas and coarse particles are also minor contributors, but they usually constitute less than 10% of extinction and are usually directly attributable to mobile sources and resuspended dust. The contribution of each source type to light extinction can

be estimated by using the CMB to estimate contribution of each source to each visibility-reducing species, multiplying this contribution by the appropriate extinction efficiency, and summing the products for each source. This approach was taken in SCENIC Denver[37] which found the major contributors to light extinction arising from primary motor vehicle exhaust, primary vegetative burning, and secondary ammonium nitrate. The lack of direct apportionment of ammonium nitrate to the emitters of its precursors (livestock for ammonia, motor vehicles and power plants for nitrogen oxides) was a shortcoming of this study. As noted above, the science has progressed to a point that such an apportionment may be feasible with the appropriate ambient measurements and estimates of transformation rates, deposition rates, and aging times.

McDade[38] describe a summertime experiment to quantify the affects of the southern California urban area on visibility in the southwest desert. Perfluorocarbon tracer gases were released from power plants in the Los Angeles Basin during the summer of 1987. Samples were taken and analyzed for these gases as well as for $PM_{2.5}$ particles, halocarbons and hydrocarbons at the Cajon Pass through which Los Angeles pollutants exit into the desert. A regional-scale profile of gaseous and particulate species is to be estimated from these measurements, and this profile will be used to apportion similar measurements made at remote sites in southern Nevada and in the Grand Canyon. The presence or absence of perfluorocarbons at the remote sites is to be used in the CMB to determine the validity of apportioning visibility extinction in the desert to pollutants from southern California.

The Winter Haze Intensive Tracer Experiment (WHITEX[39]) complements that of McDade[38] by studying the feasibility of apportioning wintertime visibility extinction in the Canyonlands and Grand Canyon National Parks to regional-scale and nearby power plant source-types. Deuterated methane tracer releases from a power plant stack were an integral part of this study with receptor measurements of $PM_{2.5}$ and PM_{10} particles and light extinction made at several receptor sites on the Colorado Plateau during the winter of 1987. Several multivariate receptor modeling and dispersion modeling approaches, including the CMB, are being applied and evaluated using the WHITEX data.

Gorse[40] compared source contributions to elemental carbon, one of the most important of the particulate species which affect visibility, derived from the CMB and PCA receptor models. The PCA attributed approximately 45%, on average, of the elemental carbon concentrations to primary motor vehicle emissions while the CMB model attributed approximately 30% of the elemental carbon to motor vehicles. The authors observed that their results contained large uncertainties because the PCA-derived source profiles appeared to co-mingle different source-types, while the available CMB motor vehicle profiles did not have elemental carbon measurements.

These applications demonstrate that the CMB is applicable to the apportionment of suspended particulate matter and light extinction. Its assumptions are known and the effects of deviations from them can be tested in general and specific cases. CMB modeling software is standardized and in common use. The software contains validation performance measures which provide quantitative estimates of model and measurement uncertainty.

References

[1]Watson, J.G., 1984a: Overview of receptor model principles. *JAPCA*, **34**, 620.

[2]Chow, J. C., 1985: *Development of a Composite Modeling Approach to Assess Air Pollution Source/Receptor Relationships.* Doctor of Science Dissertation, Harvard University, Cambridge, MA.

[3]Hopke, P.K., 1985: Receptor modeling in environmental chemistry. *Chemical Analysis,* 76, John Wiley & Sons, New York, NY.

[4]Watson, J.G., J.C. Chow, D.L. Freeman, R.T. Egami, P. Roberts and R. Countess, 1987a: *Model and Data Base Description for California's Level I PM10 Assessment Package.* DRI Document 8066-002.1D1, Draft Report, Prepared for the California Air Resources Board, Sacramento, CA.

[5]Watson, J.G., J.C. Chow and N.F. Robinson, 1987b: *Western States Acid Deposition Project Phase I: Volume 4–An Evaluation of Ambient Aerosol Chemistry in the Western United States.* Prepared for the Western States Acid Deposition Project by Systems Applications, Inc., San Rafael, CA, SYSAPP-87/064.

[6]Freeman, D.L., J.G. Watson. J.C. Chow and R.T. Egami, 1987: *Level I PM10 Assessment Package User's Manual.* Prepared for California Air Resources Board, Sacramento, CA. by Desert Research Institute, Reno, NV.

[7]Watson, J.G., 1979: *Chemical Element Balance Receptor Model Methodology for Assessing the Sources of Fine and Total Particulate Matter.* Ph.D. Dissertation, University Microfilms International, Ann Arbor, MI.

[8]Pasquill, F., 1974: *Atmospheric Diffusion.* 2nd edition, Ellis Honwood, England.

[9]Benarie, M. M., 1976: *Urban Air Pollution Modeling Without Computers*, U. S. Environmental Protection Agency Document EPA- 600/4-76-055.

[10]Seinfeld, J.H., 1975: *Air Pollution: Physical and Chemical Fundamentals.* McGraw-Hill, New York, NY.

[11]Watson, J.G, 1989: *Transactions, Receptor Models in Air Resources Management, Air and Waste Management Assoc.,* Editor, Pittsburgh, PA.

[12]Watson, J.G., J.A. Cooper and J.J. Huntzicker, 1984: The effective variance weighting for least squares calculations applied to the mass balance receptor model, *Atmos. Environ.*, **18**, 1347.

[13]Britt, H.I. and R.H. Luecke, 1973: The estimation of parameters in nonlinear, implicit models. *Technometrics*, **15**, 233.

[14]Williamson, H.J. and D.A. DuBose, 1983: Receptor model technical series, volume III: user's manual for chemical mass balance model. EPA-450/4-83-014, U.S. Environmental Protection Agency, Research Triangle Park, NC.