Source apportionment of atmospheric trace gases and particulate matter: Comparison of log-ratio and traditional approaches

M.A. ENGLE[1], J.A. MARTÍN-FERNÁNDEZ[2], N.J. GEBOY[1], R.A. OLEA[1], B. PEUCKER-EHRENBRINK[3], A. KOLKER[1], D.P. KRABBENHOFT[4], P.J. LAMOTHE[5], M.H. BOTHNER[6], and M.T. TATE[4]

[1] U.S. Geological Survey, Reston, Virginia, USA, engle@usgs.gov
[2] Dept. d'Informàtica i Matemàtica Aplicada, Universitat de Girona, Spain
[3] Dept. of Marine Chemistry and Geochemistry – Woods Hole Oceanographic Inst., Woods Hole, Massachusetts, USA
[4] U.S. Geological Survey, Middleton, Wisconsin, USA
[5] U.S. Geological Survey, Denver, Colorado, USA
[6] U.S. Geological Survey, Woods Hole, Massachusetts, USA

Abstract

In this paper we compare multivariate methods using both traditional approaches, which ignore issues of closure and provide relatively simple methods to deal with censored or missing data, and log-ratio methods to determine the sources of trace constituents in the atmosphere. The data set examined was collected from April to July 2008 at a sampling site near Woods Hole, Massachusetts, along the northeastern United States Atlantic coastline. The data set consists of trace gas mixing ratios (O_3, SO_2, NO_x, elemental mercury [Hg^o], and reactive gaseous mercury [RGM]), and concentrations of trace elements in fine (<2.5 µm) particulate matter (Al, As, Ba, Ca, Cd, Ce, Co, Cs, Fe, Ga, Hg, K, La, Mg, Mn, Na, P, Pb, Rb, Sb, Sr, Th, Ti, V, Y, and Zn) with varying percentages of censored values for each species.

The data were separated into two subcompositions: $\mathbf{s_1}$, which is comprised by RGM and particulate Hg (HgP), which are both highly censored; and $\mathbf{s_2}$ which includes all of the trace elements associated with particulate matter except Hg, and the trace gases O_3, SO_2, NO_x, and Hg^o. Principal factor analysis (PFA) was successful in determining the primary sources for constituents in $\mathbf{s_2}$ using both traditional and log-ratio approaches. Using the traditional approach, regression between factor scores and RGM and particulate Hg concentrations suggested that none of the sources identified during PFA led to positive contributions of either reactive mercury compound. This finding is counter to most conventional thinking and is likely specious, resulting from removal of censored data (up to >80% of the entire dataset) during the analysis.

Log-ratio approaches to find relationships between constituents comprising $\mathbf{s_2}$ with RGM and HgP (i.e., $\mathbf{s_1}$) focused on log-ratio correlation and regression analyses of alr-transformed data, using Al as the divisor. Regression models accounted for large fractions of the variance in concentrations of the two reactive mercury species and generally agreed with conceptualizations about the formation and behavior of these species. An analysis of independence between the subcompositions demonstrated that the behavior of the two constituents comprising $\mathbf{s_1}$ (i.e., RGM and HgP) is dependent on changes in $\mathbf{s_2}$. Our findings suggest that although problems related to closure are largely unknown or ignored in the atmospheric sciences, much insight can be gleaned from the application of log-ratio methods to atmospheric chemistry data.

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

1. Introduction

Multivariate data analysis techniques are routinely used in atmospheric science to quantify inputs and identify sources of particulate matter and trace gases in the troposphere (Thurston and Spengler, 1985). Methods applied to apportion sources of atmospheric constituents have grown substantially more complicated, allowing for assigning sample and analytical uncertainty, estimating geographic source areas, and demonstrating model uniqueness (Hsu et al., 2003; Kim et al., 2004). However, nearly all of these models and methods ignore three primary issues: 1) most atmospheric chemistry datasets used in the models typically contain a fraction of censored measurements (values below method detection limits); 2) some fraction of the values are missing (i.e., non-operational equipment, lack of sampling during dangerous conditions, power outages, etc.), and 3) data are compositional, thus classical analysis ignoring closure suffers from artifacts. Not only are these issues rarely addressed, but very few studies have examined their impact on source apportionment methods. The purpose of this paper is to provide a first comparison between traditional methods, which use typical algorithms to replace missing and censored values and ignore the constant sum constraint, and log-ratio methods specifically designed for compositional data (Atchison, 1986). These two approaches are applied to a compositional dataset of trace gas and elemental fine (<2.5 μm) particulate chemistry from a sampling site near Woods Hole, Massachusetts, along the northeastern United States Atlantic coastline, collected from April to July 2008.

2. Study Description

From 2005–2009, the U.S. Geological Survey and colleagues have collected atmospheric data to identify sources and examine cycling of atmospheric mercury in coastal environments. Data have been collected from sites along the U.S. Atlantic and Gulf of Mexico coastlines and from the island of Puerto Rico in the Caribbean. The sites range in latitude from 18.38°N (El Yunque, Puerto Rico) to 44.37°N (Acadia National Park, Maine). Deposition, sources, and characterization of atmospheric mercury at most of these sites are detailed in Engle et al. (2010). Data discussed in this paper come from one of these coastal sites, near Woods Hole, Massachusetts (Figure 1).

The dataset presented here consists of trace gas mixing ratios (O_3, SO_2, NO_x, elemental mercury [Hg^o], and reactive gaseous mercury [RGM]), and concentrations of trace elements in fine particulate matter (Al, As, Ba, Ca, Cd, Ce, Co, Cs, Fe, Ga, Hg, K, La, Mg, Mn, Na, P, Pb, Rb, Sb, Sr, Th, Ti, V, Y, and Zn) with varying percentages of censored and missing values for each species. The data were collected to identify sources of trace elements, and in particular reactive mercury (Hg) species (i.e., RGM and particulate Hg [HgP]), to the region, based on their associations and elemental profiles of emission sources. The latter species are particularly important because traditional source apportionment methods often account for <50% of their variability.

Sources of the trace elements in the particulate matter samples from the Woods Hole site were previously examined by Kolker et al. (2010) using positive matrix factorization (a data analysis approach) and concentrated-weighted trajectory analysis (an air mass trajectory approach). Four primary sources were identified in the study: 1) geogenic dust (identified by large contributions of Al, Ti, Ce, Fe, Y, Cs, Rb, Sc, and La) transported by continentally-derived air masses and air masses coming north along the coast from the Gulf of Mexico (possibly Saharan dust); 2) sea salt (identified by contributions of Na, Mg, Sr, K, and Ca) coming from the Atlantic Ocean and Hudson Bay; 3) smelters and other anthropogenic sources (characterized by large contributions of Pb, Cd, and Zn) primarily associated with air masses passing over known metal refining emission sources in New York and eastern Canada; 4) and fossil-fuel combustion (identified from large contributions of Mo, Sb, V, Ni, Cu, As, and Ba) derived from sources along the U.S. East Coast and from the Ohio and Tennessee Valleys, regions with large numbers of coal-fired power plants.

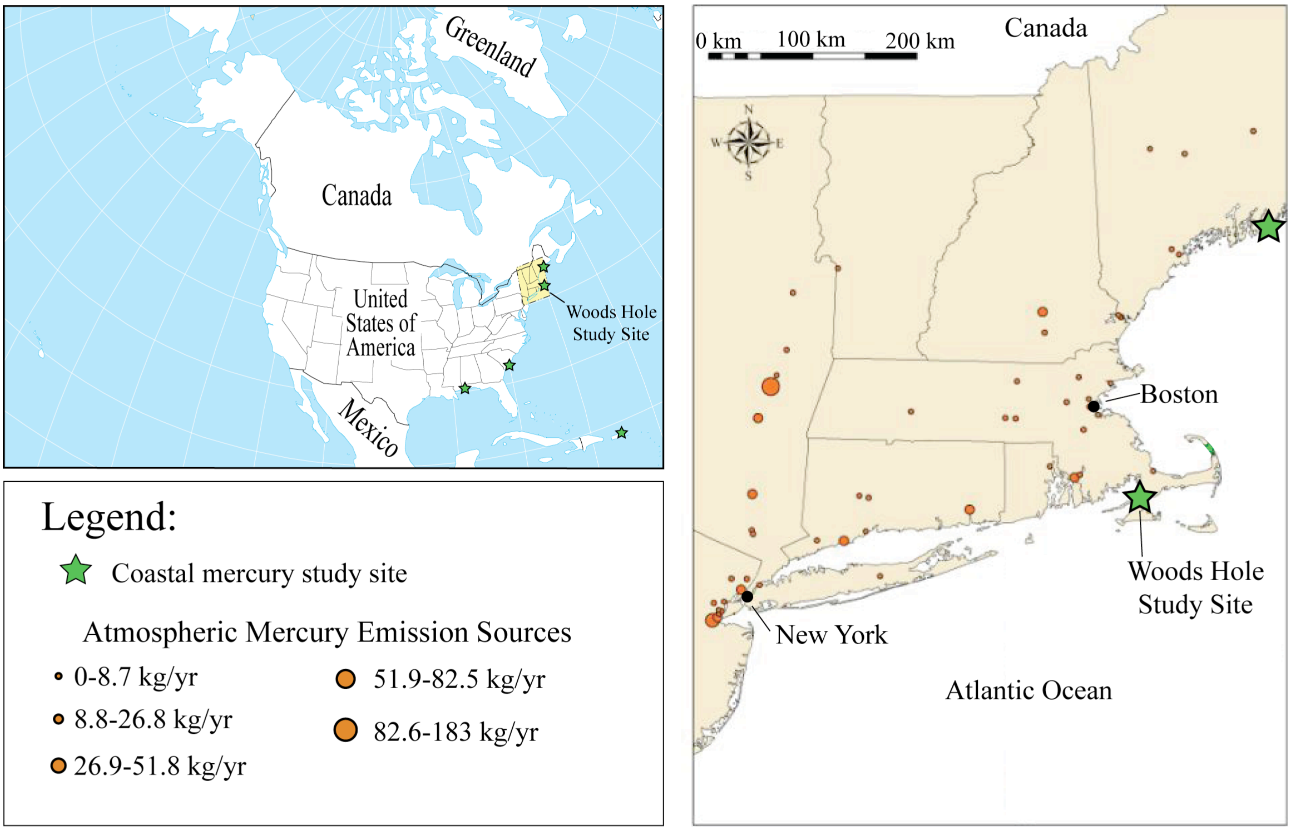

Figure 1. Maps showing location of the current study area along with study sites from previous investigations. Atmospheric mercury emission data taken from the U.S. Environmental Protection Agency Toxic Release Inventory (http://www.epa.gov/triexplorer/).

3. Methods and Data Processing

Details of sample collection and analysis are not within the scope of this paper, but are detailed in Engle et al. (2008) and Kolker et al. (2008). Because the measurements and samples represent typical conditions on different time scales (5 minutes to 2–3 days), the trace gas and mercury speciation data were averaged to correspond to the periods of collection for fine particulate matter samples; these samples exhibited the longest sampling intervals (typically 24–72 hours). After averaging, data were available for 73 events, corresponding to the periods of collection for the fine particulate matter samples. Data corresponding to the July 4th sampling event were anomalous, due to impact from fireworks, and were not examined further. Constituents (except for target constituents RGM and HgP) for which >25% of the samples were below method detection limits were excluded from examination. Of the remaining 30 constituents, 1.7% of the values were missing and 6.3% were censored. Because of the interest in the RGM and HgP results, the composition was split into a two subcompositions: a 2-part subcomposition comprised by RGM and HgP ($\mathbf{s}_1$); and a 28-part subcomposition ($\mathbf{s}_2$) containing trace elements in particulate matter, O_3, SO_2, NO_x, and Hg^o. Hereafter, we call $\mathbf{s}_1$ the *dependent* subcomposition, because we are interested in its dependence on $\mathbf{s}_2$.

3.1 Traditional Data Analysis Methods

Following more typical approaches to source apportionment of atmospheric constituents, censored values were replaced with ½ of the corresponding method detection limits and missing values were assigned estimated values calculated from the standard expectation-maximization (EM) algorithm. To group similarly behaving constituents in order to determine primary atmospheric sources, $\mathbf{s}_2$ was analyzed using minimum covariance determinant (MCD)-based robust principal factor analysis (PFA) with Varimax rotation (Reimann et al., 2008). This robust version of PFA was applied to minimize influence from outliers present in the dataset. These outliers remained in the data despite scaling and Box-Cox transformation. The number of factors to retain in the PFA model was determined from a screeplot. Several variations to the PFA model were investigated in which the number of factors and rotation algorithms were varied and robust versus non-robust covariance estimates were compared; only the most interpretable model is presented here.

Each factor from the PFA model was interpreted as a distinct atmospheric source (see results section below). To estimate the contribution of each of these sources to RGM and HgP at the site during the study period, scores from each factor were regressed, using least trimmed squares, against the species in the dependent subcomposition. Due to the large proportion of censored data for the two species in the dependent subcomposition (RGM = 40%, HgP = 72%), the regression analysis was limited to non-censored data for RGM and HgP. Although methods are available to deal with the high proportion of censored data in the dependent variables, following typical methods employed in source apportionment studies, no attempt was made to use them.

3.2 Log-ratio Data Analysis Methods

Given that a relatively small percentage of the entries in the data matrix are censored (1.7%), a multiplicative replacement of censored values was made (Martín-Fernández et al., 2003). Censored values were replaced with 65% of the corresponding method detection limits and observed values were modified so as to preserve the constant sum constraint. Missing data were imputed using two log-ratio methods: a *k*-nearest neighbor (knn) procedure and the iterative model-based imputation technique (Adj), both described in Hron et al. (2010). Both techniques provide fairly similar results, but only results using the Adj imputation will be presented here.

Similar to the methods in Section 3.1, a log-ratio approach to robust PFA of $\mathbf{s}_2$ was completed using the method of Filzmoser et al. (2009) whereby the data were mapped to a multivariate real space using an isometric log-ratio transformation (ilr) in order to calculate the MCD. Once PFA was performed, using MCD as the covariance matrix, the results were transformed into centered log-ratio (clr) space for easier interpretation. Unlike traditional methods, where each factor is generally interpreted to represent one source category (e.g., geogenic dust), results from PFA of log-ratio transformed data usually result in at least two sources for each factor (one source defined by constituents with positive loadings and the other source defined by constituents with negative loadings; Reimann et al., 2008). Because of the contribution from multiple sources to each factor, regression between factor loadings from PFA of $\mathbf{s}_2$ with a log-ratio transformation of $\mathbf{s}_1$ were less interpretable, in terms of assessing source contribution to RGM and HgP concentrations.

As an alternate method to investigate the potential sources of the two species that comprise $\mathbf{s}_1$, based on individual constituents in $\mathbf{s}_2$, a mixture of correlation, partial correlation, stepwise regression, and best-subset selection regression methods were applied to the log-ratio transformed data. First, the calculations were made on additive log-ratio (alr) transformed data, where Al was used as the divisor for the full composition. Second, a compositional analysis of independence (Aitchison, 1986) was conducted to determine the nature of independence between $\mathbf{s}_1$, $\mathbf{s}_2$, and the total sum vector of concentrations from constituents in $\mathbf{s}_1$ and $\mathbf{s}_2$ (**t**). Lastly, to examine controls of the log-variance on the two dependent variables, the log-ratio of HgP to RGM was regressed against alr-transformed data of $\mathbf{s}_2$.

4. Defining sources using principal factor analysis

The most interpretable PFA model developed using traditional methods suggests that there were six primary atmospheric sources (i.e., factors) for the constituents investigated, during the period of study (Figure 2). The high loadings of elements associated with crustal material (i.e., Al, Ce, Fe, Ti, and Y) indicate that the first source is geogenic dust. High positive loadings of Na, Mg, and Sr with negative loadings of NO_x (a combustion byproduct) likely represent input of sea salt particles, which is expected for data from a coastal site. The third factor likely represents input from metal smelters and other industrial facilities, given the high positive loadings of base metals such as Cd and Pb. Although high loadings of P are difficult to characterize, large loadings of Zn, Ba, and Sb in Factor 4 are typical of road dust, brake dust, and other vehicular-related elements (Thorpe and Harrison, 2008). Coal combustion is a major source of Hg^o, As, and Sb, while oil combustion can produce large quantities of V (Nriagu, 1989). Therefore, we attribute coal- and oil-fired power plants as the primary source of elements in Factor 5. The final source is dominated by input of O_3 with lesser contributions from SO_2, As, and Sb. These latter three constituents are tracers of coal combustion and may represent input from more distant sources than those contributing to Factor 5. This interpretation is supported by the association of O_3, a secondary pollutant that is often formed downwind of source regions, with Factor 6 and V with Factor 5; oil combustion (a major source of atmospheric V) is fairly limited in the United States and three oil-fired power plants are located within 100km of the site. Similar results were found using positive matrix factorization on the fine particulate data in Kolker et al. (2010).

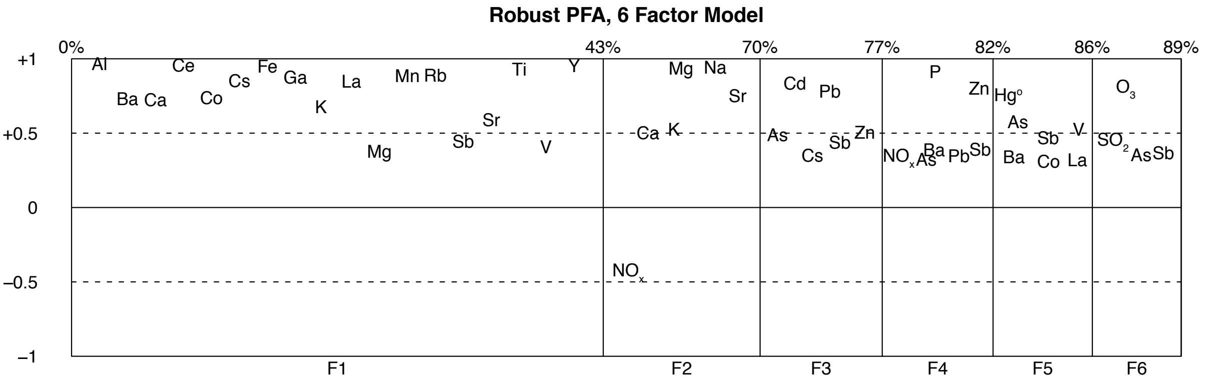

Figure 2. Plot of factor loadings for a 6-factor robust PFA of the data using traditional analysis methods. The x-axis is scaled based on the relative amount of variability explained by the 6 factors. Percentages at the top of the plot indicate the cumulative explained variability of the PFA model, from left to right. The y-axis shows loadings for each factor. Note that loadings <|0.3| are not plotted, as their contributions to the factors are minimal.

Factor loadings for PFA of clr-transformed variables comprising $\mathbf{s_2}$ (Figure 3) show similar associations between and among constituents to those observed using traditional approaches. The most obvious difference between results from the traditional approach (Figure 2) and the log-ratio methods (Figure 3) is that for the latter, each factor represents at least two sources (i.e., one producing large positive loadings and one producing large negative loadings), while in the former, each factor represents a single source. For example, positive loadings of Ce, Ti, Y, Al, and Fe in Factor 1 (Figure 3) likely reflect input of geogenic dust rich in clay minerals (similar to Factor 1 in Figure 2) while negative loadings of Cd, Pb, Zn, As, and Sb represent input from metal smelters and refineries (similar to Factor 3 in Figure 2). As the two sources sit on opposite ends of a link in the PFA biplot, Aitchison and Greenacre (2002) suggest that the relative inputs of the two sources may be controlled by a single degree of freedom. In other words, results from PFA of the clr $\mathbf{s_2}$ subcomposition indicate that each factor could represent a continuum of inverse contributions from two sources. Thus, the log-ratio PFA suggests that for Factor 1, inputs of geogenic clay minerals may be inversely related to metal refineries and smelters. Similar relationships may be interpreted for the remaining factors. One potential complication appears because multiple, geogenic-type sources are observed in the PFA model (positive loadings in Factor 1, positive loadings in Factor 3, and positive loadings in Factor 4). It is difficult to decide whether these different geogenic sources are truly different (e.g., loadings of Ce, Ti, Y, and Al in Factor 1 suggest input of clay minerals while loadings of Cs, Rb, Ca, and Al in Factor 3 represent input from felsic rocks).

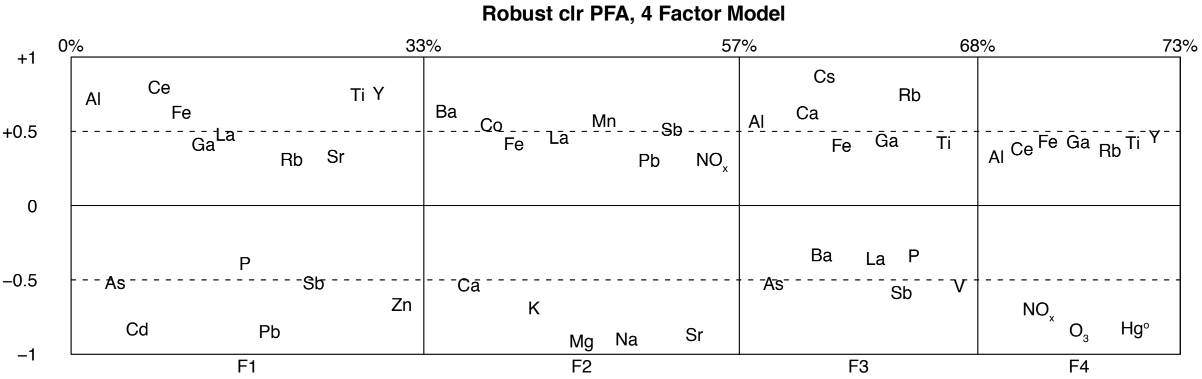

Figure 3. Plot of factor loadings for a 4-factor robust PFA of the data using log-ratio methods. The x-axis is scaled based on the relative amount of variability explained by the factors. Percentages at the top of the plot indicate the cumulative explained variability of the PFA model, from left to right. The y-axis shows loadings for each factor. Note that loadings <|0.3| are not plotted, as their contributions to the factors are minimal.

5. Examining controls on RGM and HgP

Following from the traditional approach to PFA, contributions of the sources to concentrations of HgP and RGM were estimated using a robust regression analysis between factor loadings from $\mathbf{s_2}$ and variables in $\mathbf{s_1}$. For this approach, samples in which HgP or RGM are censored were ignored for analysis of the respective compound; ignoring data heavily biases the regression results because a large fraction of HgP and RGM data are censored (Figure 4). The only factor to show a significant relationship with RGM ($r^2 = 0.38$, $p<0.01$) was Factor 5 (local coal/oil combustion), but the slope of the regression line was negative suggesting no additive contribution. No factors showed a significant relationship (at $p<0.05$) with HgP concentrations. These results indicate that despite having nearly four months of data, no obvious sources were shown to contribute HgP and RGM to the study area, using the traditional approach. We find these results unlikely as HgP and RGM are co-emitted from a variety of anthropogenic sources in the region (Kolker et al., 2010) and are likely specious due to the removal of censored data from the analysis.

Log-ratio approaches to find relationships between constituents comprising $\mathbf{s_2}$ with RGM and HgP (i.e., $\mathbf{s_1}$) focus on log-ratio correlation and regression analyses of alr-transformed data. First, an alr-transformation was applied to the full composition using Al as the divisor. Log-ratios of constituents showing a strong ($r>|0.7|$) correlation with alr-HgP include Hg^o, O_3, NO_x, Zn, and Pb. Additionally, alr-Hg^o exhibits a strong ($\rho>|0.4|$) partial correlation with alr HgP. Stepwise-regression results for alr-HgP provided the following model, which accounts for 90% of the log-ratio variance:

$\ln(HgP) = –6.72 – 0.24\ln(Ba) + 0.16\ln(Ca) + 1.0\ln(Hg^o) + 0.25\ln(Mn) + 0.17\ln(Sb) – 0.26\ln(V) + 2.08\ln(Al)$ + Error (residual standard error: 0.250).

A very similar regression model was generated from best-subset selection log-ratio regression (Bayesian Information Criterion):

$\ln(HgP) = –8.75 – 0.35\ln(Ba) + 0.25\ln(Ca) – 0.17\ln(Cd) + 0.92\ln(Hg^o) + 0.31\ln(Mn) + 0.13\ln(NO_x) + 0.28\ln(Sb) – 0.25\ln(V) + 2.12\ln(Al)$ + Error (residual standard error: 0.243),

where the adjusted R^2 is 90%. Both models suggest that elevated concentrations of HgP are associated with positive inputs of Ca, Hg^o, Sb, and Al and negative contributions of Ba and V. Associations of HgP with Hg^o and Sb indicate that HgP appears to be associated with anthropogenic sources, such as distal power generation (distal power sources are primarily coal-fired power plants and thus produce far less V than local oil-based power plants). This finding agrees with previous investigations showing the largely anthropogenic source of elevated HgP concentrations (Keeler et al., 1995).

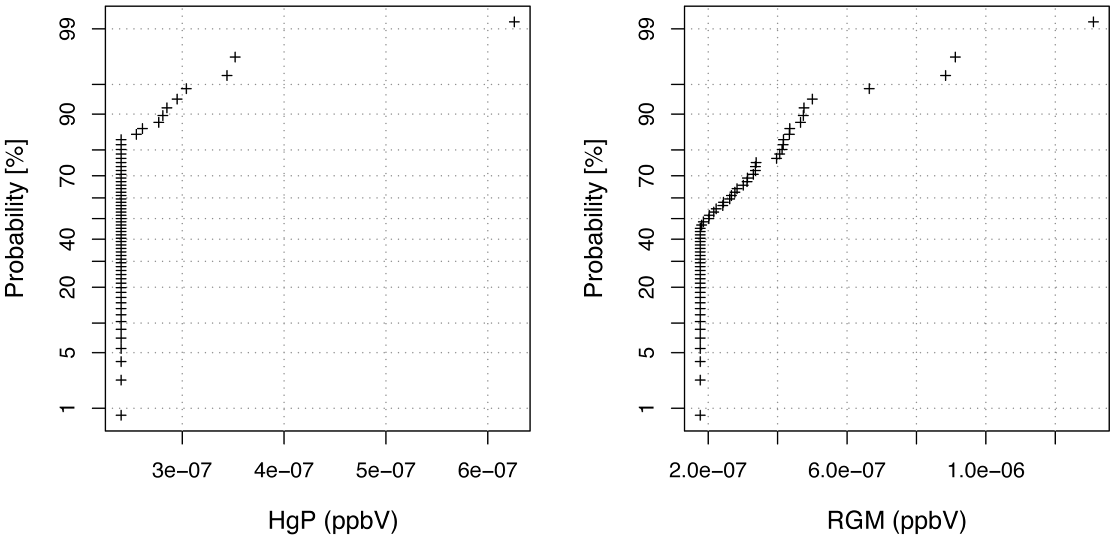

Figure 4. Concentration-probability plots for HgP and RGM. Censored data are plotted at the detection limits of their respective methods. These plots show that a large fraction of HgP and RGM data are censored.

An identical approach was taken to examine the possible RGM sources at the study site. Additive log-ratio transformed RGM correlates strongly (r>|0.7|) with log-ratios of many of the same constituents as alr-HgP: O_3, NO_x, Pb, and Hg^o. However, partial correlations are strongest (ρ>|0.4|) between alr-RGM and log-ratios of O_3, Mg, and Na. The partial log-ratio correlations are positive for O_3 and Na, and negative for Mg. Stepwise regression generated a model predicting RGM concentrations, which accounts for 83% of the variance:

$$\ln(\mathrm{RGM}) = -22.48 - 0.53\ln(\mathrm{Hg^o}) - 0.32\ln(\mathrm{K}) + 2.32\ln(\mathrm{Mg}) - 2.00\ln(\mathrm{Na}) + 1.66\ln(\mathrm{O_3}) + 0.19\ln(\mathrm{P}) - 0.11\ln(\mathrm{SO_2}) + 2.21\ln(\mathrm{Al}) + \mathrm{Error}$$ (residual standard error: 0.371).

A similar model was generated to predict RGM using a best-subset selection method:

$$\ln(\mathrm{RGM}) = -20.96 + 0.27\ln(\mathrm{Cs}) - 0.63\ln(\mathrm{Hg^o}) - 0.47\ln(\mathrm{K}) + 2.22\ln(\mathrm{Mg}) - 1.87\ln(\mathrm{Na}) + 1.71\ln(\mathrm{O_3}) + 0.18\ln(\mathrm{P}) - 0.11\ln(\mathrm{SO_2}) + 2.30\ln(\mathrm{Al}) + \mathrm{Error}$$ (residual standard error: 0.365),

where the adjusted R^2 is 84%. Both models are characterized by elevated concentrations of Mg, O_3, and Al leading to predicted high RGM concentrations, whereas low RGM concentrations are associated with Hg^o, K, Na, and SO_2. Previous work on examination of RGM sources in coastal sites, including this one, suggests that RGM is primarily formed through secondary photochemical reactions in moderately polluted, oxidizing air, rather than being emitted directly from a source (Engle et al., 2010). Secondary formation is one possible reason why traditional source apportionment methods can only account for a small percentage of RGM at some sites (i.e., the proportion that is directly emitted from an atmospheric source with other co-contaminants). The positive slopes of Mg and O_3 and negative slopes of Hg^o and SO_2 in the regression equations are consistent with this conceptual model in that Mg and O_3 are indicative of an oxidizing, coastal air mass (an ideal environment for photo-oxidation) while negative slopes of Hg^o may indicate its loss during conversion to RGM via photochemical reactions (Engle et al., 2010). A negative slope between SO_2 and RGM also indicates that RGM at the site is secondary; RGM and SO_2 are typically co-emitted from coal-fired power plants and other known RGM sources (Kolker et al., 2008). However, the positive slope for Mg and negative slope for Na is difficult to interpret given that both elements are thought to be primarily derived from the same source (i.e., sea salt aerosols; Figures 2 and 3).

Caution is highly recommended when one interprets these regression models because a simplification of non-significant coefficients has been applied. In such case, the invariance caused by permutation is violated. In other words, once a simplification of the model is considered, the invariance of the back-transformed results in relation to a change in a divisor of the alr-transformation is not ensured. Nonetheless, no relevant problems were detected in the multiple regression diagnostic plots and indices. In particular, most of the values produced by the analysis of variance inflate factors that do not appear to cause multicollinearity between the alr-transformed independent parts.

To further investigate the relationship between $\mathbf{s_1}$ and $\mathbf{s_2}$ subcompositions, an analysis of independence between these two subcompositions and the total sum vector $\mathbf{t}=(t_1, t_2)$ was conducted (Aitchison, 1986). Assuming alr-normality of the data ($\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{t}$), chi-square tests indicate that for all three, the hypothesis of independence is rejected. In particular, the hypothesis of independence with regards to neutrality on the left part ($\mathbf{s}_1$) is rejected. Results from this analysis indicate that we can assume that the behavior of the two constituents comprising $\mathbf{s_1}$ (i.e., RGM and HgP) is dependent on changes in $\mathbf{s_2}$. This finding is notable given that no major conclusions about the inputs to RGM and HgP could be drawn from the PFA results using the traditional approach.

After the analysis of independence, a final log-ratio regression model was conducted to examine the control on the log-ratio of HgP/RGM relative to the alr-transformed data of subcomposition $\mathbf{s_2}$. The stepwise regression model accounted for only 49.6% of the variance, but still provides some insights into the data:

$$\ln(\mathrm{HgP/RGM}) = 16.7 + 1.58\ln(\mathrm{Hg^o}) - 1.18\ln(\mathrm{Mg}) + 1.15\ln(\mathrm{Na}) - 1.52\ln(\mathrm{O_3}) - 0.17\ln(\mathrm{P}) + 0.14\ln(\mathrm{Al}) + \text{Error (residual standard error: 0.471)}$$

The model suggests that of the two reactive Hg species, HgP is likely co-emitted and transported with $\mathrm{Hg^o}$ to the site, while RGM tends to be dominant in oxidizing (e.g., O_3-rich) conditions. However, the negative contribution from Mg and the positive contribution from Na is difficult to interpret as both species are typically attributed to inputs from sea salt aerosols (Figures 2 and 3).

Again one must be careful when interpreting such models because in all of them a simplification of non-significant coefficients implies that invariance by permutation is violated. In other words, once a simplification of the model is considered, the invariance of the back-transformed results in relation to a change in a divisor of the alr-transformation is not ensured. These kinds of problems, also detected in Barceló-Vidal et al. (2011) for time series modeling, are not exclusive to the alr-transformation. For example, Barceló-Vidal et al. (2011) show that when ilr-linear models are applied and simplified, the invariance by changes of basis is violated.

6. Summary of Findings

This paper provides a first comparison of traditional and log-ratio methods to identify sources of trace constituents measured in both gas and particle phases during a 4-month field campaign at a site along the northeastern United States Atlantic coastline. Despite using relatively simple methods for dealing with censored and missing data and ignoring the effects of closure, results from PFA analysis using traditional methods showed similar results to those for log-ratio methods. However, application of the traditional methods to assess contributions from sources to concentrations of HgP and RGM were largely unsuccessful. By comparison, the log-ratio approach was promising in creating regression models that accounted for large fractions of the variance in concentrations of the two reactive mercury species. The regression models generally agreed with conceptualizations about the formation and behavior of these species. However, inclusion of elements typically associated with disparate sources made interpretation tricky. Results from analysis of independence between subcompositions demonstrated that the behavior of the two constituents comprising $\mathbf{s_1}$ (i.e., RGM and HgP) is dependent on changes in $\mathbf{s_2}$. These findings suggest that although problems related to closure are largely unknown or ignored in the atmospheric sciences, much insight can be gleaned from the application of log-ratio methods to atmospheric chemistry data. Therefore, our efforts must be focused to construct an appropriate multiple linear regression model using log-ratios. The log-ratio regression models may be improved through the definition and utilization of chemically meaningful sequential binary partitions between the 30 constituents (Egozcue and Pawlowsky-Glahn, 2005). Once such an ilr-basis is generated, a multiple regression model would be formulated between the balance of dependent constituents against the balances or log-contrast formed with the constituents in the $\mathbf{s_2}$ subcomposition

Acknowledgements

We acknowledge support of the USGS Toxic Substances Hydrology Program, the USGS Energy Resources Program, the National Science Foundation's Small Grant for Exploratory Research Program, and for initial support, the USGS Mendenhall Postdoctoral Program. This research has also been supported by the Spanish Ministry of Science and Innovation under the projects "CODA-RSS" (Ref. MTM2009-13272) and by the Agµencia de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya (Ref. 2009SGR424).

The authors wish to thank Carles Barceló-Vidal (Universitat de Girona) and Kevin Jones (U.S. Geological Survey) for their constructive criticisms and comments on an earlier version of the paper. Michael Casso (USGS Woods Hole Science Center) and Mark Olson (USGS Wisconsin Water Science Center) assisted in operation of the mobile mercury lab.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. and M. Greenacre (2002). Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51 (4),375–392.

Barceló-Vidal, C., Aguilar, L., and Martín-Fernández, J.A. (2011). Compositional VARIMA time series. In Pawlowsky-Glahn and Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications.* John Wiley & Sons.

Egozcue, J. and V. Pawlowsky-Glahn (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology* 37 (7), 795–828.

Engle, M.A., M.T. Tate, D.P. Krabbenhoft, A. Kolker, M.L. Olson, E.S. Edgerton, J.F. DeWild, and A.K. McPherson (2008). Characterization and cycling of atmospheric mercury along the central US Gulf Coast. *Applied Geochemistry* 23 (3), 419–437.

Engle, M.A., M.T. Tate, D.P. Krabbenhoft, J.J. Schauer, A. Kolker, J.B. Shanley, and M.H. Bothner (2010). Comparison of atmospheric mercury speciation and deposition at nine sites across central and eastern North America. *Journal of Geophysical Research* 115 (D18), D18306, 13 p.

Filzmoser, P., K. Hron, C. Reimann, and R. Garrett (2009). Robust factor analysis for compositional data. *Computers & Geosciences* 35 (9), 1854–1861.

Hron, K., M. Templ, and P. Filzmoser (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics and Data Analysis* 54 (12), 3095–3107.

Hsu, Y., T.M. Holsen, and P.K. Hopke (2003). Comparison of hybrid receptor models to locate PCB sources in Chicago. *Atmospheric Environment* 37 (4), 545–562.

Keeler, G., G. Glinsorn, and N. Pirrone (1995). Particulate mercury in the atmosphere: Its significance, transport, transformation and sources. *Water, Air, & Soil Pollution* 80, 159–168.

Kim E., P.K. Hopke, T.V. Larson, and D.S. Covert (2004). Analysis of ambient particle size distributions using Unmix and Positive Matrix Factorization. *Environmental Science & Technology* 38, 202–209.

Kolker A., M.A. Engle, W.H. Orem, J.E. Bunnell, H.E. Lerch, D.P. Krabbenhoft, M.L. Olson, and J.D. McCord (2008). Mercury, trace elements and organic constituents in atmospheric fine particulate matter, Shenandoah National Park, Virginia, USA: A Combined Approach to Sampling and Analysis. *Geostandards and Geoanalytical Research* 32 (3), 279–293.

Kolker A., M.A. Engle, B. Peucker-Ehrenbrink, M. Bothner, D.P. Krabbenhoft, N.J. Geboy, and P. Lamothe (2010). Geochemistry of atmospheric aerosols on Cape Cod, MA: Implications for air quality and human health. *Geological Society of America Abstracts with Programs* 42 (5), 353.

Martín-Fernández J.A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology* 35 (3), 253–278.

Nriagu, J. (1989). A global assessment of natural sources of atmospheric trace metals. *Nature* 338, 47–49.

Reimann, C., P. Filzmoser, R. Garrett, and R. Dutter (2008). Statistical Data Analysis Explained: Applied Environmental Statistics with R. Wiley & Sons. 362 pages.

Thorpe, A. and R.M. Harrison (2008). Sources and properties of non-exhaust particulate matter from road traffic: A review. *Science of The Total Environment* 400 (1–3), 270–282.

Thurston, G.D. and J.D. Spengler (1985). A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. *Atmospheric Environment* 19 (1), 9–25.