APPENDIX F:

HOW THE SOURCE APPORTIONMENT TOOLS WORK

Appendix F: How the Source Apportionment Tools Work

F.1 General Set-up

All of the receptor modeling tools, UNMIX, PMF, and SAS, assume that each of the significant sources to the observed aerosol concentrations have fixed (or nearly constant) profiles. That is, if we could observe just one source at a time, then even though the total mass from the given source would change from day to day, the relative amount of each species is constant. So, for example, the $PM_{2.5}$ from cigarette smoke is mostly organic carbon, but it is also about half a percent potassium and a quarter of a percent chlorine. Wood burning, on the other hand, has nearly equal amounts of potassium and chlorine where both are over half a percent of the total mass. Now the consistency of these may seem questionable, but consider the wood sources. For an area, the available wood types are fairly fixed and wood smoke will only be observable if many people are burning wood. Hence, only an area-averaged wood smoke profile will be observed and this will be fairly constant. A similar averaging effect takes place for other sources as well.

The above understanding of what the receptor sees leads to a mathematical model for the observed aerosol. Up to measurement error, each observation, of say potassium, is the sum of the contributions from the various sources. Moreover, each contribution is the product of the total mass from the corresponding source and the relative percent of that source that is potassium. The same will be true for chlorine (and all the other measured species). So,

The i^{th} mass measurement =
(i^{th} amount of mass from source 1) + (i^{th} amount of mass from source 2) +
(i^{th} amount of mass from source 3) + …

The i^{th} potassium measurement =
(i^{th} amount of mass from source 1)(the ratio of potassium to mass in source 1) +
(i^{th} amount of mass from source 2) (the ratio of potassium to mass in source 2) +
(i^{th} amount of mass from source 3) (the ratio of potassium to mass in source 3) + ...

The object is to find non-negative estimates for each of the values that show up in parentheses above. The tools do this using quite different techniques.

F.1.1 PMF

PMF essentially finds the set of non-negative estimates that minimize the total squared differences between each side of the above equalities added up over all days and all species. Because the absolute values and relative errors of the mass measurements and the potassium values are very different, each of the squared differences is weighted with an estimate of the uncertainty of the measurement. Hence, the squared quantities considered are not squares of mass difference between the estimate and the measured value, but rather the number of times greater than the measurement uncertainty the difference represents.

The PMF results are usually very precise compared to UNMIX, but can be biased. The main reason is that the user determines the number of sources in the solution. The program is then forced to find the best possible fit based on an incorrect assumption. Another problem is that "best fit" just is not the same as truth. The result is geared to mathematically best explain the observed concentrations within the model assumptions, but if the model assumptions are not quite met then the solution may be biased. In fact, the program includes the Fpeak setting to "unbias" the solution produced.

F.1.2 UNMIX

UNMIX obtains a solution based on evidence within the data that support the number of sources. UNMIX's solution is based on the additional assumption that there are periods within the data that each source makes no significant contribution to the receptor's observed concentrations. Each time this happens there is one fewer source contributing to the mixture, and, hence, there is a reduction in the dimension/complexity of the possible mixture. Consider the Figures 3.1.

The program first searches for the "edges" in the data created when a source is "turned off." The algebraic formula for these edges is uniquely linked to the profiles. So, when the edges are found, the program solves for the profiles and then uses an ordinary least squares regression to get the relative strengths. Finally, the solutions are checked for non-negativity. If the solutions fail the non-negativity filtering criteria, then the program reports back "No Feasible Solution Found." There are filtering parameters that the user can set that control how sensitive to negative values these criteria are. The "No Feasible Solution Found" message will also occur if the program fails to find enough edges in the data.

F.1.3 SAS's NLMIXED

SAS's "Proc NLMIXED" is a very general set of algorithms. It will try to estimate the parameters for any mathematically expressible relationship of *observed value y = a function with unknown parameters + error*. The "NL" part of the name refers to the fact that it fits non-linear functions. The "MIXED" part of the name refers to the type of situation that we have with SA. In addition to the random measurement error, part of the functional relationship is random. Specifically, the contributions change randomly from day to day. The convergence can be slow and it is not assured. However, unlike PMF or UNMIX, SAS has several numerical routines that can be used to do the estimation. One may work better than another.

The SAS procedure, as it has been used for SA, models the concentration with a function that is basically the one shown in the introduction to this Appendix. The procedure works better when the parameters can be scaled into similar ranges. Hence, like PMF, the function works with difference between the modeled and the measured values divided by the laboratory uncertainty. This weighted difference is then treated as a random variable that we model with a normal distribution.

The consequence of the above modeling framework is that the observations are viewed as the outcome of a random process with a distribution that can be mathematically described (or at

least numerically approximated) in terms of the parameters of interest. Hence, in principle, for any given set of parameters there is a fixed probability of observing data that matches the data that were observed. More importantly, this can be turned around to find the set of parameters that maximize this probability. This is a general technique called maximum likelihood estimation.

Consider a much simpler example. Suppose an "unfair" coin is flipped 10 times and results in heads 9 of those 10 times. It is fairly easy in this case to describe the probability distribution associated with flipping an unfair coin 10 times. The only parameter in this case is the probability of getting heads on any one flip. Now ask the question, "What is the value for this parameter that maximizes the probability of getting 9 out of 10 heads?" One can do a lot of math to show that the obvious answer, 0.9, is the most likely one. In other words, it is much more likely that a coin that averages 9 out of 10 heads will yield 9 out of 10 heads on a particular set of flips than that a fair coin that averages only 5 heads out of 10 would yield heads 9 out of 10 times.

In practice, the general technique of maximum likelihood estimation is just a matter of finding the parameters that maximize a particular object function that is constructed from the modeling assumptions about the data. The details of both the object function and the maximization process are handled by the routines in SAS. It is more important to know that maximum likelihood estimation has near optimal standard errors for the estimated parameters. SAS automatically computes these with the parameter estimates. For large samples, as are needed for SA, reliable confidence intervals can also be generated for each parameter estimate and are part of the default SAS output.