Section 7. Data Transformation Policy and Guidance

Variations in $PM_{2.5}$ measurements attributed to methodological differences should be minimized to support consistent data analysis across temporal and spatial regimes. For example, it would be erroneous to infer that 20% of the $PM_{2.5}$ measured in an urban area is due to local sources based on a comparison of the concentrations measured by monitors in an urban area to concentrations from upwind sites, if the instrumentation at the upwind sites are biased low by 20%, relative to the instrumentation used in the urban area. Realistically, $PM_{2.5}$ measurements should "look" like measurements taken by an FRM. This is because of the richness of the available FRM data base and due to the difficulty in ascribing a "reference" check for aerosol measurements.

Non-FRM samplers generally operate at a higher temporal resolution than FRMs and many will operate where there is no FRM, thus helping to fill spatial gaps in the FRM network. However, to be able to use data from multiple types of $PM_{2.5}$ mass monitoring networks (FRM, non-FRM) in the same analysis, the data must be comparable. Comparable means that if the various samplers were spatially and temporally interchanged, approximately the same concentrations would be measured. To achieve comparability, it is possible to transform, using statistical models, non-FRM data to look like FRM data or vice versa. Due to the interest in FRM-like concentration surfaces, the remainder of this section will only address transforming data from non-FRM samplers to produce FRM-like measurements.

Due to the inherent differences in measurement principles between FRM and PM continuous monitors there may be biases between the measurements obtained from an FRM and continuous monitor. If the bias is consistent through time and across space, a standardized correction factor could be used to produce FRM-like measurements from the continuous monitors. However, since mass concentration and composition and environmental conditions vary, a standard correction may not be practical on a national scale but may be achievable on a more regional scale. This section provides information about the development of transformations to produce FRM-like measurements from continuous measurements.

Based on preliminary analyses summarized in Section 2, developing a statistical model to relate concentrations from continuous samplers (predominantly TEOMs) to FRM samplers is achievable, although the complexity of the model varies by location and may vary through time. The complexity likely is a function of the stability of the composition of the aerosol, the stability of the meteorology (temperature and humidity), and the continuous monitoring methodology. The following guidance for developing transformations is based on the experience gained in analyzing the limited collocated FRM/continuous database to date. The database is limited due to temporal representativeness (at best 2 years since the FRM network was deployed in 1999), spatial representativeness (continuous samplers have been and continue to be deployed predominantly in large urban areas), and non-FRM sampling techniques. The database is predominantly based on data reported to AIRS. Prior to 2000, it was not possible to determine whether the data from a continuous monitor was reported after being adjusted by "correction" factors. Beginning in 2000, AIRS method codes were expanded so that it would be possible to determine whether correction factors had been applied, although it is not possible to specify the form or parameter estimates of the adjustment. These new method codes appear not to be accurate for all sites, as seen in Section 2, making it a further challenge to determine appropriate transformations.

A balance between forcing a particular measurement principle to mimic another (i.e., the FRM) is a significant complication that must be recognized in this task. The practical needs for data analysts demand some level of comparability. However, there is intrinsic value in the very differences that emerge between measurement systems due to the complex character of aerosols. The intention clearly is not to define the FRM as truth, but rather to recognize the practicality of the existing network. These considerations of basic measurement principles are embodied in this transformation guidance. Where relationships between two measurement systems exhibit simple linear and constant character, one can probably assume the difference in measurement approach does not result in a significantly different indicator of ambient aerosol. Such simple relationships are the foundation for accommodating REMs that can be compared to the NAAQS. On the other hand, more complex relationships between a candidate system and the FRM suggest that a significantly different aerosol property is being accounted for (likely varies over time or space) in one system relative to the other. This does not mean one system is superior to the other, but reasonable judgement suggests a limit to forcing a system to mimic the FRM for regulatory use, but to accommodate the system for other data uses within the limits of data comparability guidelines. This latter approach reflects the concept underlying the expanded use of CACs.

The guidance on transformations will be broken into two sections, one for the CAC and one for the REM. The guidance for acceptable transformations for REMs will be very strict and limited to simple transformation models. Acceptable transformations for CACs will be less strict. For either case, recall that the performance criteria presented in Section 6 is based on the transformed continuous measurements. Note that if the performance criteria are met with the raw continuous measurements, then no transformation is required. That is, transformations need not always be developed.

Regardless of whether the continuous sampler is a CAC or REM, measurements should be reported to AIRS. Given that data users might not understand the differences in the sampling methodologies, it is recommended that the data be entered AFTER applying a transformation to produce FRM-like measurements. However, it will be important for other data uses to know what transformations have been applied. EPA will be investigating possible ways to include the transformation information in AIRS so that it will be possible to “back out” the transformation and have the original, non-FRM measurements.

Transformation Guidance for CAC

Even though the data from a CAC will not be used for direct comparison to the NAAQS, they should meet the DQOs, as described in Section 6. Although this is not a requirement, it is strongly recommended for comparability of measurements across the network. The data used for evaluation in the DQO process are those that have been transformed to be FRM-like; that is, the DQOs are not necessarily based on the raw data from the non-FRMs. This section describes the process for developing the transformations for CACs. The rationale and details for the selection of many of these criteria are included in the EPA document *Reporting an Air Quality Index (AQI) Using Continuous $PM_{2.5}$ Data: Data Quality Objectives (DQOs) and Model Development for Relating Federal Reference Method (FRM) and Continuous $PM_{2.5}$ Measurements (Attachment C).*

Step 1. Create daily non-FRM measurements. If the non-FRM data are collected more frequently than daily, the sub-daily intervals should be averaged before comparing to the FRM data. At least 75% of the sub-daily intervals should be valid to consider the average to be valid. Also, the sub-daily intervals to be averaged should be those that most closely span midnight to midnight, the operating interval of the FRMs.

Step 2. Determine if there are sufficient data to develop statistical model. The model to relate the non-FRM and FRM data should be based on data from all four seasons and have at least 104 valid pairs of data, approximately evenly distributed through each season. It is recommended that each season have at least 20 valid pairs. If there are not more than 100 valid pairs approximately evenly distributed through the seasons, it is recommended that additional data be collected. The 100 pairs need not be from only one year.

3. Develop a statistical model. The statistical model relating the non-FRM and FRM data should have the FRM data as the response variable (also called the dependent variable) and minimally must include the non-FRM measurements from *Step 1* as an explanatory (independent variable). The number and type of explanatory variable allowed in unlimited. The model can be based on the data as is or can be based on the natural logarithms of the data. The final R^2 between the measured and predicted FRM measurements should be 0.80 or greater.

4. Spatial extent for use of one transformation. Section 8 describes the process for determining the area within which one transformation may be used for all of the continuous samplers, regardless of whether the continuous sampler has been previously collocated with an FRM.

5. On-going evaluation of transformation and its spatial extent. The statistical model should be revisited every 3 years, or more frequently if there is reason to believe a change in the relationship between the non-FRM and FRM may have occurred. Possible reasons for such changes include, but are not limited to, a change in sampling methodology, change in aerosol composition due to control strategies, or different meteorological regimes than what was observed during the development of the statistical model. If a new statistical model is more appropriate, that model should be used from that date forward. That is, one model would be used up to one date and the next model would be used for subsequent dates.

Transformation Guidance for REM

The data from REMs must meet the DQOs, as described in Section 6. The data used for evaluation in the DQO process are those that have been transformed to be FRM-like; that is, the DQOs are not necessarily based on the raw data from the non-FRMs. This section describes the requirements for the transformations. Because the data are intended to be used for NAAQS comparisons, the allowable statistical models and parameter estimation will be explicitly defined. The reason for this specificity is to ensure that two independent data analysts will produce the same transformation and hence will produce the same FRM-like concentrations. Most of the details of this guidance are unknown at this time, due to limited data, but the issues that need to be addressed and a time line for addressing them is included. The guidance components are as follows.

Step 1. Manipulation of non-FRM data, prior to development of statistical model. This section will detail how data are aggregated to produce a daily number to be used to compare to the FRM. Issues to address will include handling of missing data, producing averaging periods that are approximately midnight to midnight, and handling of negative or zero concentrations prior to aggregation. Data completeness will also be addressed. Likely, at least 75% of the sub-daily intervals should be valid to consider the average to be valid.

Step 2. Identification of pairs to use in development of transformation model. This section will address the number of required valid pairs, temporal representativeness of those pairs (e.g., whether highest and lowest seasons are sufficient or if every season must be represented), range of concentrations spanned by the pairs (need a good spread so that the model is appropriate and can be used for prediction through wide range of concentrations), handling of negative or zero concentrations, handling of concentrations less than some cutoff value (e.g., minimum detection limit), identifying and handling influential pairs.

Step 3. Development of statistical model. This section will detail how the statistical model relating FRM and aggregate non-FRM collocated data is to be developed. The only model allowed will be one for which the aggregate non-FRM data is the only explanatory variable and the FRM data is the response variable, that is, only a slope and intercept will need to be estimated. Issues include whether the raw or natural logarithm of the raw data are to be modeled, the required R^2 between the measured and predicted FRM measurements, the equations for estimating the slope, intercept, and R^2 especially if seasons are not equally represented in the data set (that is, should the estimates be weighted).

Step 4. Inferences to be drawn from the statistical model. This section will discuss how to determine the spatial representativeness of the model (what is the area that can use the same transformation, which will be discussed in Section 8) and the temporal representativeness of the model (for how long is the model valid). When a transform is found to be no longer appropriate, what is done with the previously transformed data? Is the old transformed used until up to one date and then the new transform used for subsequent dates?

Step 5. On-going evaluation of statistical model and its spatial extent. At least 30% (rounding up) of the non-FRM sites must be permanently collocated with FRMs to provide the data needed to evaluate regularly the reasonableness and consistency of the transforms. The collocated sites should be distributed to represent different composition and meteorological regimes. Issues to cover include the frequency at which the transformation is formally evaluated to determine whether it is still appropriate. For example, it would not be practical to have a transformation that is changed every month or quarter, but the transformation should be reviewed at some frequency.

Due to the numerous issues to developing statistical models, EPA will establish a panel to recommend solutions to the various issues listed above. The panel will be comprised of people conversant in statistics, ambient air monitoring, and air quality management. Solutions to these issues and final guidance on the development of transformations for regulatory data use are expected to be completed by the end of calendar year 2003.