APPENDIX A:

DATA MANIPULATION

Appendix A: Data Manipulation

This appendix chronicles the key elements in the data manipulation and preparation for the study. The concentration data used were either from an AQS pull provided by EPA or from the NY DEC website. EPA also provided the HYSPLIT output for each site. The protocol for generating these data is also described.

A-1. Determining Uncertainties and Filling in Missing and Below MDL Data

The AQS data includes 3 sites (sites 250250042, 360050083, and 482011039) that ran co-located speciated $PM_{2.5}$ monitors from the same vendor/manufacturer (a part of the Mini-Trends study). Generally these monitors ran from February 2000 through July 2000. These co-located data were used to establish a key portion of the uncertainties. The goal was to establish a base “uncertainty” for each species. Ideally, these uncertainties would be the standard error of the measurements. However, the goal only needs to establish a relative weighting for each species.

First, the data from all three sites were combined. In this way, the uncertainties could be averaged across sites and manufacturers. For each species, the standard deviation of the paired differences was found. These values are all comparable to the maximum MDL for each species. (The different machines use different volumes and, hence, have different MDLs.) Since the MDLs represent a lab uncertainty, the uncertainties for each above-MDL measurement for a species was based on the average of the maximum MDL for the species and the standard deviation of the paired differences. This provides a relative weighting of the species. The standard deviation of paired differences, however, should be larger than the standard error of the measurements by a factor of square root of two. Hence, the final step was to divide the relative weighting factors by the square root of two. This yielded a standard error for each species (see Table A-1).

In the PMF modeling, the standard error described above was used for values greater than the MDL. The above procedure only applies to measured values above the MDL. For values below the MDL, MDL/2 was substituted, and the uncertainty was set to equal two times the uncertainty calculated above. If a data value was missing, then it was filled in with the site-species mean times the ratio of the $PM_{2.5}$ measurement for that day divided by the site’s mean $PM_{2.5}$. In these cases, the uncertainty used was two times the sum of the filled in value and the species uncertainty.

Finally, the uncertainty from the $PM_{2.5}$ measurement (based on the nylon filter of the speciation sampler) was used for the uncertainty of any co-located FRM value.

Table A-1. Species' Uncertainties Prior to Adjusting for Below MDL or Missing Data

Species	Uncertainty ($\mu g/m^3$)	Species	Uncertainty ($\mu g/m^3$)
FRM	0.8419	Nickel	0.0070
PM2.5	0.8419	Nitrate	0.0716
Aluminum	0.0273	Organic Carbon	0.4111
Ammonium	0.1169	Potassium	0.0063
Arsenic	0.0012	Potassium Ion	0.0090
Barium	0.0291	Selenium	0.0010
Bromine	0.0010	Silicon	0.0456
Calcium	0.0188	Sodium	0.0434
Chlorine	0.0158	Sulfate	0.2641
Chromium	0.0009	Sulfur	0.0355
Copper	0.0016	Tantalum	0.0105
Elemental Carbon	0.1357	Tin	0.0088
Iron	0.0207	Titanium	0.0022
Lead	0.0027	Vanadium	0.0008
Manganese	0.0014	Zinc	0.0012

A-2. Use of the FRM Data

For all sites except the Houston site, co-located FRM mass measurements were made. Hence, for each site except Houston, there are two independent measurements of the $PM_{2.5}$ concentration for each sampling event: one from the FRM instrument and one from the speciation sampler. Generally, these measurements were simply used as an additional species (which effectively weighted the mass two times any other species except the sulfur/sulfate pair). Since these measurements should always be equal (disregarding measurement error), any missing values were filled in with the other whenever needed. Moreover, in these cases, the uncertainty for the inputted value was doubled. In the case of the Houston site, all of the "FRM" values were filled in with the mass measurement value from the speciation sampler for consistency with the other results. In the source apportionment output, the average of these apportioned values was used for the mass of the source.

A-3. Plan for Running HYSPLIT4 Trajectory Model

The following sections are from EPA's plan for running the HYSPLIT4 model. This was used as a protocol document for the actual implementation.

A-3.1. Potential Uses of a Back Trajectory Database

The PM fine Analysis team intended to run the HYSPLIT4 trajectory program for four trends sites for the entire 2000 calendar year. It was concluded that a back trajectory database would be helpful for analyses to determine potential source origins of elevated ambient air quality. The trajectory database can be used as input to residence time analyses, and in combination with wind roses to help locate the probable sources for the high PM fine values that were measured during the year 2000.

A-3.2. Trajectory Program

The intended model to calculate trajectories is called HYSPLIT (Hybrid Single-Particle Lagrangian Integrated Trajectory). It is a complete system for calculating air parcel trajectories, dispersion, and deposition simulations. This model was developed by the National Oceanic and Atmospheric Administration (NOAA). It was upgraded in April 2001 as a joint effort between NOAA and Australia's Bureau of Meteorology. The new version includes improved advection algorithms, updated stability and dispersion equations, a new graphical user interface, and the option to include modules for chemical transformations.

Other available trajectory models are ATAD, CAPITA Monte Carlo model, and a hierarchical modeling approach. The hierarchical modeling method is not well suited for this problem since it is still under development and has not been completed. HYSPLIT and the Monte Carlo model have been compared and were found to produce very similar results when treated as an ensemble. HYSPLIT has also been compared with observable data such as smoke and dust clouds, and good agreement was found. A paper describing a comparison made between HYSPLIT and the CAPITA Monte Carlo model trajectories was written by Bret Schichtel and Paul Wishinski. A brief report on this comparison is located at the following internet address: http://capita.wustl.edu/otag/Reports/vtdecair/ vtdecair.html#TrajectoryLength. Although some differences were noted between trajectories of the two models, when the results were applied in ensemble techniques, the differences appeared relatively minor. HYSPLIT is probably the most used and most scrutinized of the models. The Monte Carlo model and ATAD require time to QA the input data. The input data for HYSPLIT have already been preprocessed and QA'ed by NCEP/NCAR. Therefore, the model that appears to be most suited to this project is HYSPLIT, mainly because it has good input data that are readily available.

The HYSPLIT model can be run interactively on the Web through the READY system on the NOAA site, or the code executable and meteorological data can be downloaded to a Win95/98/NT PC. The Web version has some limitations to avoid saturation of the web server, but the PC version is complete. Meteorological data files must be downloaded for the PC version. HYSPLIT has been extensively run for many years by numerous agencies and many comparisons have been performed. Each model run is relatively simple to set up, execute, and plot and usually only takes a few minutes to run on the READY system that NOAA implemented on the internet site. It is also possible to plot numerous trajectories on the same map in order to do a cluster type analysis. This can only be performed on the PC, though.

HYSPLIT can do multiple simultaneous trajectories starting at the same time. Also, the user can define up to three starting height levels. Typical starting heights are 500 m, 1,000 m, and 1,500 m[1]. Computations can be performed forward or backward in time. The default vertical motion uses the omega field, consisting of u, v, and w velocity components that are

[1] Only the trajectories stating from 500 m were used in the analysis.

output from the meteorological models. Other options available are isentropic, isosigma, isobaric, and isopycnic surfaces.

HYSPLIT also has the ability to use input from multiple nested meteorological data grids. Trajectory calculations can also use archived or forecast data from meteorological models. The data conversion programs available for the trajectory model are for GRIB and netCDF file formats. Model graphics consist of Tcl/Tk GUI with on-line help, and the graphics are displayed as postscript files.

The mixing height along the path of the trajectories is another issue that was considered. This would involve tracking the mixing height and tracking the trajectory height along the trajectory path. Knowledge of the mixing height along the trajectory path could be beneficial because it could help determine whether an air mass was coming from within the mixed layer or the upper air. This information could be helpful in analyzing trajectories by placing more emphasis on trajectories below the mixing height when determining potential source locations. The newest version of the Hysplit model has the calculated mixing heights along the trajectory paths. The documentation for this version of HYSPLIT is not available to the public at this time. Roland Draxler from the National Oceanic and Atmospheric Administration is the contact for this model and updates.

A-3.3. Data Sources

Meteorological Inputs to HYSPLIT

The meteorological input fields that HYSPLIT uses are required to be in “ARL packed” format. All of the gridded meteorological data available on the READY site are in the required form and can be transferred to a PC without conversion.

The National Centers for Environmental Prediction (NCEP), which is part of the National Weather Service, execute a series of meteorological models on an operational basis primarily for forecast purposes, but the output is archived for additional purposes. Part of the Global Data Assimilation System (GDAS) is EDAS (Eta Data Assimilation System), which covers the U.S. NOAA’s Air Resources Laboratory (ARL) archives NCEP model output for air quality transport and dispersion modeling. The EDAS and GDAS data are archived using a 1-b packing method. The archives contain basic meteorological parameters such as the u- and v-wind components, temperature, and humidity. The archives differ in terms of the horizontal and vertical resolution, and in the specific fields.

The 3-hourly archive data used by the HYSPLIT model are available from NCEP’s EDAS. The EDAS originated from the operational early Eta model runs during 1995. It is an intermittent assimilation consisting of successive 3-hour Eta model forecasts consisting of 38 levels and a 48 km horizontal resolution grid with optimum interpolation. A 6-hour forecast GDAS is used as an initialized field to start the assimilation at 12 hours prior to the model start time.

Three-hour analysis updates allow for the use of high frequency observations, such as wind profiler, NEXRAD, and aircraft data. The 48 km data are then interpolated to a 40 km, Lambert Conformal Grid, which covers the continental United States.

ARL processes these data by running an archive program that extracts every other grid point of the 3-hourly, 40 km grid that results in an 80 km dataset on pressure surfaces. Some of the grid points at the edge of the model domain are removed to reduce the size of the semi-monthly files. These data are then saved to tape and shipped to the National Climatic Data Center (NCDC). The EDAS data are also put on-line at ARL's web site.

The EDAS archive data file consists of data in a synoptic time sequence without any missing time periods. Missing data are represented by nulls and the forecast hour is set to negative 1. Each file contains about two weeks of data. The first file for the month consists of days 1 through 15 and the second file contains data for the 16th day through the end of the month.

The EDAS data are on a 79 by 55 Lambert Conformal grid. The reference latitude is 35 N and the reference longitude is 95 W.

The archived data in EDAS contain a subset of the fields that are normally produced by the NCEP. The archived fields were selected according to what is most relevant for transport and dispersion studies and disk space limitations. The EDAS files are in a format compatible with the HYSPLIT model so that no conversion is necessary to run the model using the archived data base.

Originally, it was thought that the EDAS data would be the best input data set to be used for this analysis. However, it was pointed out that the FNL (Final GDAS run) data set or the NCEP re-analysis data set could also be utilized. The FNL data set consists of meteorological model output at 191 km resolution and includes late arriving conventional and satellite data observations that are not available in the EDAS data set. Also, the FNL data are more complete and have much less missing data than the EDAS data. An analysis was done to compare trajectories that were run using EDAS, FNL, and NCEP reanalysis data. The final results are not yet available. However, based on the initial results, FNL was used whenever the EDAS data were incomplete.

Air Quality

In order to perform other analyses using the trajectory output, it may be necessary to gather speciated data from the trends sites. The trends network is operated by EPA and consists of 54 monitoring locations throughout the country. They measure 59 chemical species that are 2.5 microns or less in size. These trends sites operate continuously over a 24-hour period, and the filter samples are collected every third day. Then the sampling starts again for another 24-hour period.

Of the 54 trend sites, three sites were chosen with the highest PM2.5 readings over the year 2000. Backward trajectory analyses will be performed at the eight trend sites over a 24-hour period every third day to match the data sampling.

A-3.4. Model Set-up Parameters

The HYSPLIT model will be run using EDAS data for the year 2000. It will be run using backward trajectories over a 72-hour period. The vertical motion of a trajectory will be determined using the meteorological model vertical velocity. Other user options that are required to run the HYSPLIT model are the starting time and date, and location (latitude, longitude) of a trajectory.

Since the measured PM fine data were saved every three days, the program will be run on approximately 122 days of the year. Since the data reflect a 24-hour average, trajectories will be run for 4 start times (3AM, 9AM, 3PM, and 9PM). Trajectories for every 6-hour interval within that 24-hour period will more accurately depict the sources of PM over the measured 24-hour period than a single trajectory time period. This will result in 488 trajectories for each PM monitoring site. There is an option to run forward or backward trajectories. The plan is to run backward trajectories for 5 days, but only use the first 3 days in most instances. This length of time was considered the most appropriate because a shorter run may miss the actual source location and, with a longer run, the accuracy of the trajectories would become too uncertain. Also, sensitivity testing was done with the trajectory duration. As they reduced the trajectory duration from an initial 5 days to 2 days, they found relatively minor effects on the ensemble results from 5 days down to 2.5 days. They found that the shorter durations did affect the ensemble results. This study is summarized at the following internet address: http://capita.wust1.edu/otag/Reports/vtdecair/vtdecair.html#TrajectoryLength. HYSPLIT allows for three options for vertical motion in calculating trajectories. These are model vertical velocity, isobaric, and isentropic. Modeled vertical velocity will be used and is recommended for this type of application.

The trajectory starting heights are important because they can help determine short range sources from longer range sources and the trajectories can vary widely within the mixed layer compared to above it. It has been suggested that starting heights of trajectories should be within the mixed layer, but high enough to minimize surface effects. Also, mixing heights can vary dramatically from day and night. At night, the mixing height is driven by mechanical mixing (winds) but, during the day, the mixing height is usually driven by energy from the sun. Since the solar angle is very different during the different seasons at the mid latitudes, the daytime mixing heights can vary significantly with the seasons. The stability of the atmosphere also plays a role in determining mixing heights, so the mixing heights can be different during the more stable Fall season compared to the more unstable Spring season. We are considering using seasonal and possibly diurnally varying mixing heights as a basis for deciding what start heights to use for the trajectories. The trajectory model allows for three different starting heights for trajectories at a time. Initially, starting heights were to be 500 m., 1,000 m, and 1,500 m. above ground level. The 500 m height is considered the most representative starting height for trajectory calculations, but in order to determine the best trajectory calculation to fit the PM fine data, the other starting heights were going to be used in addition to the 500 m height.

A-3.5. Terrain Heights

The model-defined terrain heights in the EDAS data sets that are used in the trajectory calculations have been smoothed out by the modeling process. This smoothing will be taken into account and the trajectory heights will be adjusted to get the actual terrain heights. This adjustment is necessary because the meteorological models tend to smooth out valleys and mountains due to sampling at selected points on a grid. The terrain adjustment can be done using the vertical sounding program that is available on the web in the READY system. The text listing of the vertical sounding shows the mean sea-level pressure (MSLP) and the surface pressure (SPRS). An example site is given below that demonstrates how the trajectory start heights will be adjusted for this task:

> Given: A station has 1,030 mb MSLP and 939 mb SPRS.
>
> Problem: Determine the appropriate input height.
>
> Solution: The difference between these two readings is 91 mb. If a hydrostatic atmosphere is assumed, then 10 meters per mb is an approximation. This gives 910 MASL, which is what the model sees as the terrain height. Since it is known that the terrain is at 951 MASL, the trajectory start height should be adjusted by adding 41 MAGL (951-910 m) if one is interested in where the air originated at some time in the past.

A-3.6. Output Data Availability

Trajectories will be performed for every third day starting at heights of 500 m, 1,000 m, and 1,500 m, and will be done at 4 start times per day. The start heights will be adjusted for the errors in modeled terrain as explained above. Backward trajectories will be run for 72 hours. Trajectories will be run for the four chosen trends stations (stations with the high PM fine measurements). Endpoints files and plot files will be archived in EPA/OAQPS's G:\user\share directory. These endpoints files can then be accessed by others and can be plotted.

APPENDIX B:

MATCHING THE SOURCE APPORTIONMENT OUTPUT TO SPECIATE PROFILES

Appendix B: Matching the Source Apportionment Output to Speciate Profiles

This appendix describes the algorithm for matching the source apportionment output profile to the profiles in the Speciate database.

The output from PMF gives the mean species mass at the receptor from each source. Hence, each value in the profile has units of $\mu g/m^3$. The profiles in the Speciate database consist of unitless ratios, which result from a measured concentration divided by a total mass concentration. There are several choices for the denominator: the total $PM_{2.5}$ from the source, the total of the species measured, or a reconstructed mass. The first is probably the ideal, but may not have been measured when the emission profile was measured since it requires a different measurement technique than the techniques for measuring the species. (Even if it was measured, it would need to have been done consistently with the current $PM_{2.5}$ measurements to "adjust" the water and semi-volatile content consistently.) The other choices for the denominator depend on the particular species that were measured, and these are not consistent from profile to profile in the database. Consequently, the matching algorithm needs to treat the profiles in a manner that recognizes these inconsistencies.

There is no point to trying to convert the source apportionment output to unitless ratios, since they are not consistent. In fact, since the range of the percent composition is so great across species, this is not a desirable scale. A more even scale is obtained by expressing the apportioned species mass as a fraction of the mean species mass for the receptor. Outlined below is how this scale is achieved for both the source apportionment output and the Speciate profiles.

The profile-matching algorithm is based on a regression. For a fixed output profile, let f_{sp} be the mean species mass apportioned to the source. Then, if a Speciate profile corresponds to the same source, there should be some "total mass value" m_p such that

$$f_{sp} = m_p \cdot f_{p,sp}$$

where

p = is the Speciate profile number (just an index); and
$f_{p,sp}$ = is the species fraction for profile number p.

The profiles considered must have at least 16 species in common with the output profile excluding ammonium, sulfate, or nitrate. So there are at least 16 equations in the single unknown. The equations are not weighted evenly. In particular, they are weighted with the mean species receptor mass. Hence, the m_p value minimizes

$$\sum_{\text{species}} \left(\frac{f_{sp} - m_p \cdot f_{p,sp}}{\text{mean}_{sp}} \right)^2$$

where mean_{sp} is the mean species mass at the receptor.

The measurement for how well the Speciate profile fits the output profile is then given by the fit score:

$$100 \cdot \sqrt{\frac{1}{N} \cdot \left[\sum_{\text{species}} \left(\frac{f_{sp} - m_p \cdot f_{p,sp}}{\text{mean}_{sp}} \right)^2 \right]}$$

where N is the number of species in common with both profiles.

SAS's weighted regression routines can quickly regress each output profile against each Speciate profile and automatically output both the mass term m_p and the quantity under the square root above. The best fitting profiles are then obtained by sorting the results by the fit scores.

Additional details

The emission profiles do not generally include the "secondary" species and, if they do, they could be quite misleading since these can be dramatically increased during transport from the source to the receptor. Also, the IC measurements are rarely included in the profiles and, consequently, are excluded. Hence, ammonium, elemental carbon, nitrate, organic carbon, potassium ion, sodium ion, sulfate, and sulfur are not considered in the automated routine for matching the profiles (but, of course, are considered in the assignment of the preliminary identifications).

Also, good fit scores can occur if the species that are in common between two profiles only include the species to which the source does not significantly contribute. To help alleviate that phenomenon, if a source contributes at least 20 percent of the receptor mass of some species, then that species must be included in the Speciate profile to be considered. This also makes sure that the "important" species are considered. Typically, when this mismatch in species occurs, the mass, m_p, associated with the Speciate profile is very different from the apportioned mass. Hence, in the graphical output, no Speciate profile is plotted if none of the 10 best fitting profiles have a value of m_p within a factor of 10 of the apportioned mass. Further, the Speciate profiles plotted in Appendix C minimize the absolute difference between the apportioned mass and m_p.