THE ANALYSIS ENGINE: A NEW TOOL FOR MODEL EVALUATION, SENSITIVITY AND UNCERTAINTY ANALYSIS, AND MORE

Alison M. Eyth*, Prashant P. Pai
UNC-Chapel Hill, Carolina Environmental Program, Chapel Hill, NC
e-mail: eyth@unc.edu, prapai@unc.edu
Web address: http://www.cep.unc.edu
Voice: (919) 966-2134   Fax: (919) 843-3113

1. INTRODUCTION

The Analysis Engine is a tool for generating tables and plots from ASCII data files. It has three main components: the table application, the plotting engine, and the statistics package. The table application provides the user interface for importing data files into the system and displays the data as rows and columns. Each data file is shown in its own tab. The table application can sort, filter, and format the data. It accesses the statistics package to compute statistics on the data, and it passes subsets of the data to the plotting engine, which creates many types of plots.

A unique feature of the analysis engine is that it allows the user to save a configuration of the table along with a set of plots. This configuration can be reloaded at a later time and applied to data files with the same format. This allows the user to quickly reproduce a table and set of plots in the same form as they were in a previous analysis engine session. This is very useful when the same analyses need to be repeated with new data sets, as is the case for model evaluation and sensitivity and uncertainty studies.

The Analysis Engine is a Java application that can run on Windows, Macintoshes, Linux, and other versions of Unix. It is provided as part of the U.S. EPA's Multimedia Integrated Modeling System (MIMS), but does not require any other parts of MIMS to run and can be used as a standalone system. For more information on MIMS, see http://www.epa.gov/asmdnerl/mims/.

2. SYSTEM COMPONENTS

2.1 The Table Application

The table application provides the interface for loading data into the analysis engine. It can currently read the following types of files: comma separated value (.csv), tab delimited, custom delimited, fixed-width columns, SMOKE report, ARFF files (used by WEKA), and several other specialized formats.

In the table application, data are presented in rows and columns, as shown in Figure 5. When multiple files are loaded, each file is shown in its own tab. The data in a column can be sorted up or down by clicking on the column header. Operations available from the toolbar include sorting using multiple data columns, showing the top or bottom *N* rows (based on data from a specific column), filtering out rows (based on the data attributes of multiple columns), hiding columns, formatting columns, creating plots, computing statistics (e.g., mean, sum, histogram, percentiles), and viewing the "analysis configuration."

The analysis configuration consists of the table configuration and the plots you have created during your session. Configurations can be saved and applied to data sets in future sessions, thereby allowing an analysis to be quickly repeated on a new data set. After the table is made to appear as the user wants it, the contents of the table can be exported to .csv, .arff, and delimited file formats. Eventually, other formats such as RTF and HTML may be supported.

2.2 The Plotting Engine

The plotting engine consists of a Java interface to the statistical package R (see http://www.r-project.org/). The plot icon on the Analysis Engine toolbar provides users with access to the plotting engine. The dialog that appears when this icon is pressed is shown in Figure 1. Programmers can access the plotting engine directly via its application programming interface (API).

The following types of plots can be created by the Analysis Engine: bar, box and whisker, cumulative density function (CDF), discrete category, histogram, rank order, XY, scatter, line,

* *Corresponding author address:* Alison M. Eyth, Carolina Environmental Program, UNC-Chapel Hill, 600 Bank of America CB#6116, Chapel Hill, NC 27599-6116

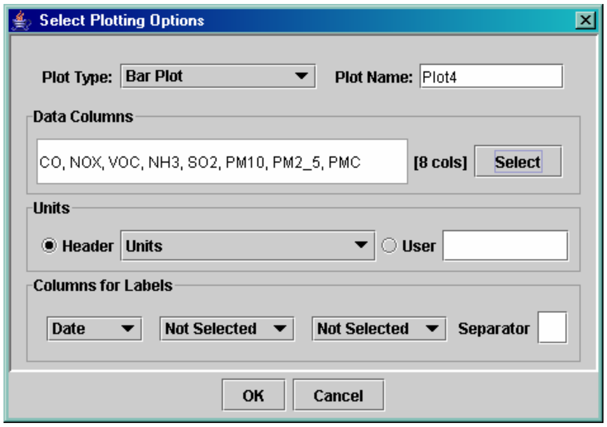
time series, regression, and tornado. The plots can be shown on the screen and saved to a variety of formats, including PDF, Postscript, JPEG, PNG, and LaTeX.

Some atypical plots are the rank order and discrete category plots. Rank order plots (Figure 2) rank the data elements in ascending or descending order and then plot them according to their rank. This plot type is useful for screening analyses. The discrete category plot (Figure 3) is similar to a bar plot in that it has names or “categories” on the X axis and numeric values on the Y axis. However, instead of showing the data values with bars, they are plotted using symbols. This makes it easy to show many data sets on a single plot.

2.3 *Statistics Package*

When the statistics icon on the Analysis Engine toolbar is pressed, the table application opens a window (see Figure 6) that allows the user to choose from a set of statistics that can be computed, along with a set of data columns for which they should be computed. Currently, basic statistics, histograms, and percentiles are available. The selected statistics are computed for each data column that was selected. After statistics are computed, the statistics for each column are shown in new data tabs within the statistics window. Basic statistics, percentiles, and histograms each appear as tables in separate tabs. Currently, all statistics are computed using the Colt package (see http://www-itg.lbl.gov/~hoschek/colt/).

The available basic statistics are minimum, maximum, sum, mean, median, standard deviation, skew, and kurtosis. If percentiles or histograms are selected, additional configuration information can be specified in the statistics window. For example, breakpoints can be specified for the histogram, and the desired percentiles can be specified. Shortcuts are provided in order to make this process as painless as possible. An example of a table that was created by computing percentiles is shown in Figure 7.



Select Plotting Options

Plot Type: Bar Plot
Plot Name: Plot4

Data Columns
CO, NOX, VOC, NH3, SO2, PM10, PM2_5, PMC [8 cols]

Units
● Header Units ○ User

Columns for Labels
Date
Not Selected
Not Selected
Separator

Figure 1. Plot Selection Dialog

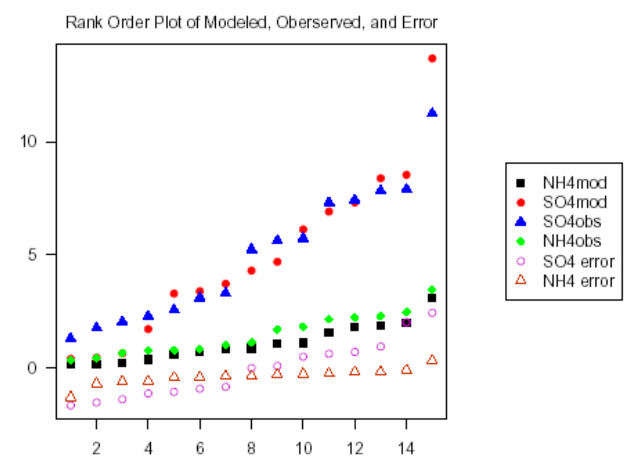


Figure 2. Rank Order Plot of Modeled, Observed, and Error Values.

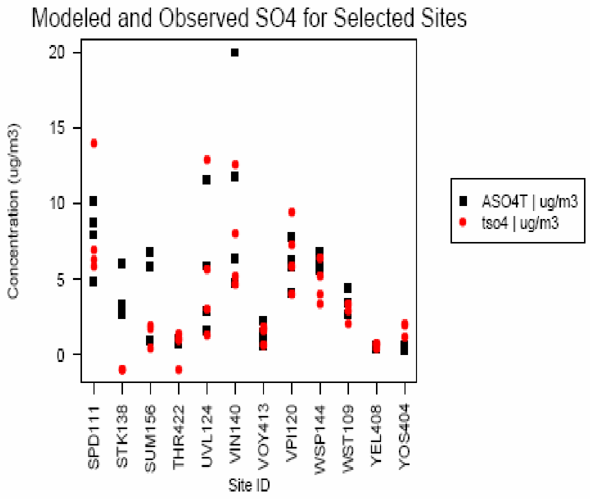


Figure 3. Discrete Category Plot of Modeled and Observed SO_4 for Selected Observation Sites.

3. APPLICATIONS FOR THE ANALYSIS ENGINE

The table and plotting tools available from the Analysis Engine can efficiently support many types of analyses, such as model evaluation, sensitivity and uncertainty analysis, and general data analysis. Using the table application, data can be sorted, filtered, and formatted into a desired state. Then the resulting data sets can be plotted and analyzed with statistics. Figures 2 and 3 are example plots from a model evaluation study. The rank order plot in Figure 2 allows for modeled, observed, and error values of several species to be examined in a summary fashion. Figure 3 shows the comparison of modeled and observed SO_4 for a filtered set of sites.

A feature of the table application that is very useful for model evaluation and other types of data analysis is the ability to save the configuration of the table and the plots that were generated to a file called an analysis configuration, discussed in Section 2.1. When an analysis configuration is loaded for a new data set, the same set of plots can be quickly regenerated. This feature can also be accessed from the command line for generating plots in an automated fashion.

Some special capabilities that support sensitivity and uncertainty analysis have been added to the Analysis Engine. One of these is the ability to compute and plot CDFs and regression lines. An example of a regression plot is shown in Figure 4. Other useful features for sensitivity and uncertainty analysis are new statistical analyses for obtaining tables of regression coefficients and correlation coefficients. These analyses will be implemented by interfacing with the software WEKA (see http://sourceforge.net/projects/weka/).

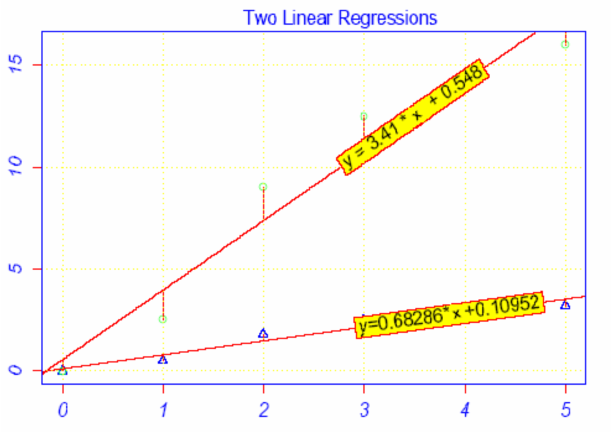


Figure 4. An Example Regression Plot.

4. FUTURE DIRECTIONS

Some of the enhancements that would be useful to implement in the Analysis Engine table application are:

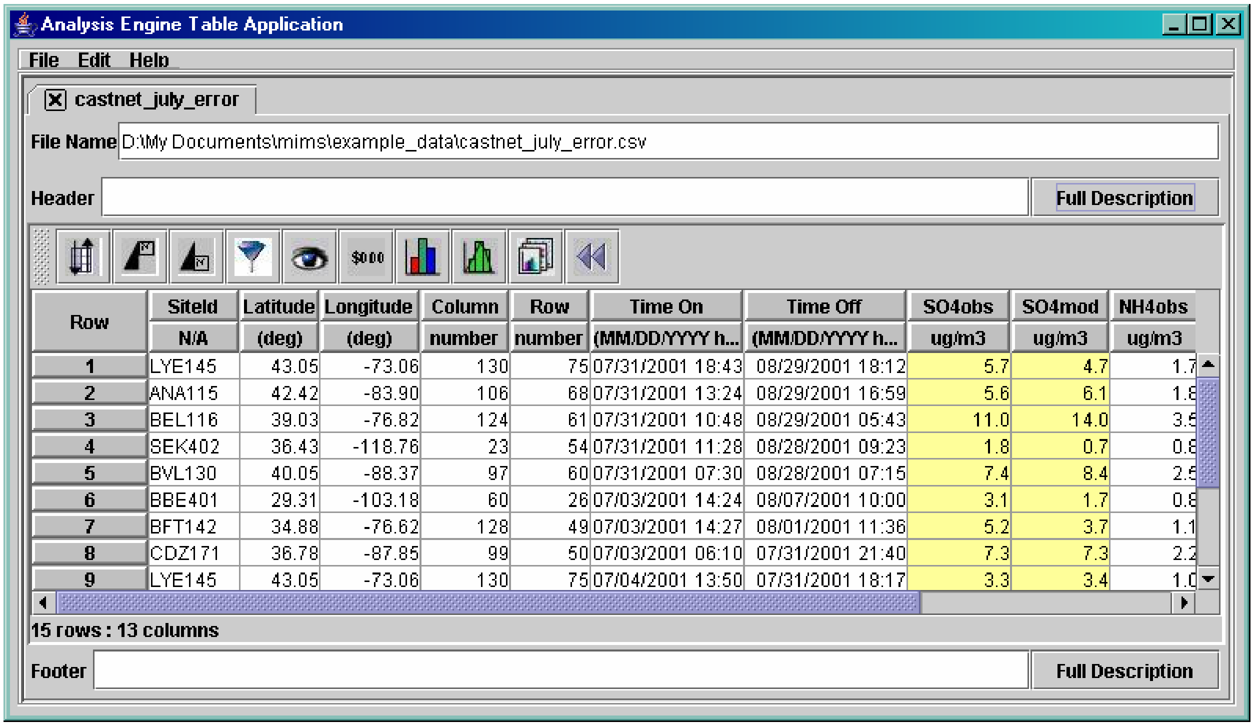
- adding rows and columns to tables that are the result of functions (e.g., difference, sum);
- creating plots and tables from data obtained from multiple tabs / files;
- copying data onto the clipboard so that they can be pasted into Microsoft Excel or other applications; and
- optionally hiding the header, footer, and file name on each tab.

An up-to-date list of the features that might be added can be found on the requests for enhancement (RFEs) page for the MIMS project in SourceForge at https://sourceforge.net/projects/mimsfw. To see the RFEs specific to the Analysis Engine, click on the RFE link in the list of links under the project title bar, then select Analysis Engine as the Category on the RFE Web page and click on the Browse button.

The project that is currently funding the development of the Analysis Engine at the Carolina Environmental Program will be completed on October 31, 2004. Some development work on the plotting engine will continue at EPA's Visualization Laboratory in FY2005. Funding for further development of the table application is dependent upon support from programs that want to use the Analysis Engine. For example, the Analysis Engine may become part of an emissions quality assurance tool.

5. ACKNOWLEDGMENTS

The development of the table application, statistics package, and GUI for the plotting engine was supported by the Multimedia Integrated Modeling System (MIMS) Modeling Support project under GSA Task Order ID 4TCG68025312. The remainder of the plotting engine was developed by Mr. Thomas Cathey at the EPA Scientific Visualization Laboratory in Research Triangle Park, North Carolina. The lead architect for the original design team for this software was Dr. Steven Fine, currently of NOAA.



File Name D:\My Documents\mims\example_data\castnet_july_error.csv

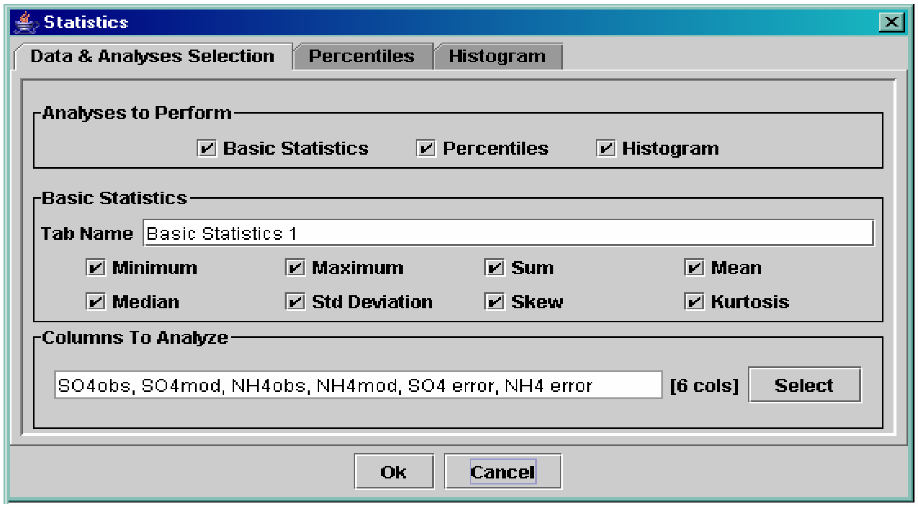
Header

Row	SiteId	Latitude	Longitude	Column	Row	Time On	Time Off	SO4obs	SO4mod	NH4obs
	N/A	(deg)	(deg)	number	number	(MM/DD/YYYY h...	(MM/DD/YYYY h...	ug/m3	ug/m3	ug/m3
1	LYE145	43.05	-73.06	130	75	07/31/2001 18:43	08/29/2001 18:12	5.7	4.7	1.7
2	ANA115	42.42	-83.90	106	68	07/31/2001 13:24	08/29/2001 16:59	5.6	6.1	1.8
3	BEL116	39.03	-76.82	124	61	07/31/2001 10:48	08/29/2001 05:43	11.0	14.0	3.5
4	SEK402	36.43	-118.76	23	54	07/31/2001 11:28	08/28/2001 09:23	1.8	0.7	0.8
5	BVL130	40.05	-88.37	97	60	07/31/2001 07:30	08/28/2001 07:15	7.4	8.4	2.5
6	BBE401	29.31	-103.18	60	26	07/03/2001 14:24	08/07/2001 10:00	3.1	1.7	0.8
7	BFT142	34.88	-76.62	128	49	07/03/2001 14:27	08/01/2001 11:36	5.2	3.7	1.1
8	CDZ171	36.78	-87.85	99	50	07/03/2001 06:10	07/31/2001 21:40	7.3	7.3	2.2
9	LYE145	43.05	-73.06	130	75	07/04/2001 13:50	07/31/2001 18:17	3.3	3.4	1.0

15 rows : 13 columns

Footer

Figure 5. Analysis Engine Table Application showing sites with the top 15 values for observed SO_4.



Analyses to Perform

☑ Basic Statistics ☑ Percentiles ☑ Histogram

Basic Statistics

Tab Name Basic Statistics 1

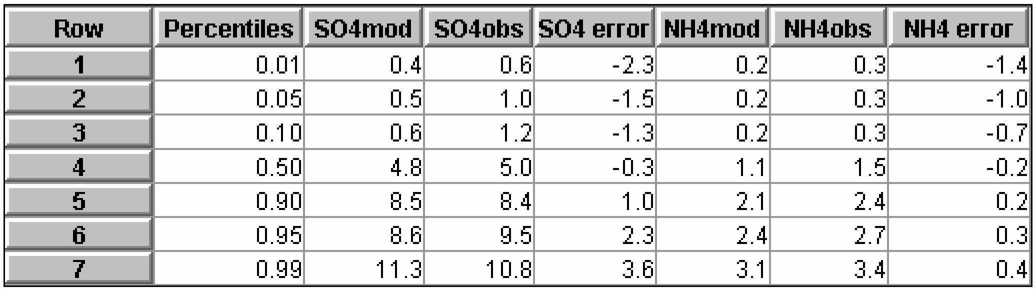
☑ Minimum ☑ Maximum ☑ Sum ☑ Mean

☑ Median ☑ Std Deviation ☑ Skew ☑ Kurtosis

Columns To Analyze

SO4obs, SO4mod, NH4obs, NH4mod, SO4 error, NH4 error [6 cols]

Figure 6. The Statistics Window.



Row	Percentiles	SO4mod	SO4obs	SO4 error	NH4mod	NH4obs	NH4 error
1	0.01	0.4	0.6	-2.3	0.2	0.3	-1.4
2	0.05	0.5	1.0	-1.5	0.2	0.3	-1.0
3	0.10	0.6	1.2	-1.3	0.2	0.3	-0.7
4	0.50	4.8	5.0	-0.3	1.1	1.5	-0.2
5	0.90	8.5	8.4	1.0	2.1	2.4	0.2
6	0.95	8.6	9.5	2.3	2.4	2.7	0.3
7	0.99	11.3	10.8	3.6	3.1	3.4	0.4

Figure 7 . A Screen Shot of a Table of Percentiles Generated by the Analysis Engine.