

**Estimating Missing Data and Analyzing
Variability of FACE Experimental Data, 1998-2001**
Detailed Report

Prepared for
Kathryn Lenz, Project Sponsor

Prepared by
Mariah Olson

July 10, 2002

Estimating Missing Data and Analyzing Variability of FACE Experimental Data

INTRODUCTION

The Forest Atmosphere Carbon Transfer and Storage (FACTS-II) Aspen Free-Air Carbon Dioxide Enrichment (FACE) Experiment studies the effects of increased levels of ozone and carbon dioxide on the ecosystems of northern hardwood forests (Dickson et al., 2000). Established in 1997, the project collects data from the US Forest Service FACE site located in Rhinelander, Wisconsin. The data is then analyzed by a number of different research groups, including a group from the University of Minnesota Duluth (UMD). Among other uses, the collected data is used in a computer simulation program being developed at UMD named ECOPHYS (Lenz and Stech, 2000; Dickson et al., 2000). My research project has required me to analyze and fill in gaps within the data collected from the FACE research site and communicate my findings to the ECOPHYS research group at UMD. I was advised by Dr. George Host, NRRI, about the biological aspects of the FACE experiment. I was also advised by Professor Kathryn Lenz and Professor Harlan Stech, Department of Mathematics and Statistics, in doing graphical and comparative analyses of the FACE data. Another undergraduate research student, Kyle Roskoski, also assisted with preparing the FACE data files and briefing me on how to effectively use the Excel spreadsheet.

ECOPHYS relies on a number of different input parameters to simulate the growth of poplar and aspen trees, including the amount of carbon dioxide or ozone in the air. Environmental inputs also include hourly readings of light, temperature, and relative humidity levels. The data collected from the FACE experiment provides the basis for

these inputs but it is often the case that recorded data is missing for several hours, or even several weeks during a growing season. The ECOPHYS model requires completed data sets to perform extensive simulation studies of the FACE site tree growth. Other FACE researchers may also be interested in working with completed data sets. Therefore, missing environmental data must be approximated. Values used to fill in the missing data can impact the ECOPHYS simulation output. For this reason, it is important that these values be close to reality. Due to high frequency fluctuations and variability among field measurements, filling in the missing data properly is an intricate and difficult procedure.

The FACE experiment is comprised of four different gas treatments, carbon dioxide only, ozone only, a combination of carbon dioxide and ozone, and a control treatment with no added gasses. Each treatment has three replications, for a total of twelve experiment rings. Missing data is often a result of equipment failure, lightning strikes, or sensor malfunctioning, which do not always affect every treatment and replication. Because of this, certain replicates will have missing data, while other replicates may have the data. Often, it seems reasonable to fill a section of missing data with the data from another replicate or treatment. However, difficulties with this procedure arise when every replicate is missing a specific section of data. Even when data from other replicates is present, filling in the data is a tedious procedure.

COMPLETED RESEARCH

I began my research project by familiarizing myself with the FACE Experiment. First, I studied a publication (Dickson et al., 2000) and the FACE website (<http://www.nrri.umn.edu/factsii>) describing a number of the experimental design details

and then I began to look at the original data files containing experimental data. I received all of the data files from Kyle Roskoski (Roskoski, 2002), a student researcher also working with the ECOPHYS research group. However, the original data was received from two researchers at the FACE site, Dr. Jaak Sober and Dr. Warren Heilman. The quantity of information contained in the data files was enormous, with an average of about 12,000 individual pieces of data for each of the twelve rings. Through the use of graphs, I was able to see the large percentage of data missing in various data files (Figure 1). Once I was familiar with the basics of the FACE Experiment, I started filling in experimental data files for the year 1998. In the absence of a reliable data model for predicting missing values, I filled in the data using two different methods. First, I checked to see if data was available from other treatments or replications during the same time period. If this was the case, data was filled on an hourly basis using the existing data values which most nearly matched the patchy file during times when data for both were present. To do this, two or more data files were compared on an hour by hour basis to see whether the files matched closely. I was also able to use hourly graphs of the experimental rings being compared as a determining factor (Figure 2). Then, the missing data was filled with the appropriate values. For light, temperature, and relative humidity variables, it was often the case that a geographically close ring was a suitable match for filling missing data. When all of the rings were missing data during the same time period, a different method had to be used. The gaps were then filled using data from the same ring, but from different yet similar days. To determine if two days within the same replication were similar, I compared data surrounding the missing section with data from other days. When a time period was found to match, the missing data was filled.

Eventually, the data files for the years 1998, 1999, and 2000 were completed in this manner.

I visited the FACE research site in Rhinelander, Wisconsin, on April 22, 2002, where I presented my project to a number of FACE researchers. I also accompanied my research group on a tour of the site, where we viewed and walked through several of the experimental treatment rings. The trees were still dormant and the gases were not yet turned on. On the tour, I saw the instrumentation used to apply the gases and to collect data for the extensive study of carbon dioxide and ozone effects on growth.

RESEARCH CHALLENGES

Analyzing the data and filling in missing data in the very large FACE data files was tedious and time consuming. I learned that graphs were an effective way to analyze the data to determine visually such things as whether experimental replications matched closely and where large portions of data were missing. I was also able to observe trends in the experimental data to see events such as when the concentrations of carbon dioxide and ozone were the greatest as well as trends in temperature, light, and relative humidity during the growing season. I also learned that graphs are excellent for summarizing my project findings and showing other FACE researchers issues with the data sets, such as large portions of missing data. Throughout the project I expanded my knowledge of Microsoft Excel by working with the large data files. Specifically, I learned how to perform many keyboard shortcuts and use Macros within the program in order to repeat a series of functions numerous times.

A number of problems arose during the process of filling in missing data. First, discrepancies between two sets of data from the same ring, one set from Dr. Jaak Sober

and another from Dr. Warren Heilman, made filling in the missing data a challenge because I did not know which data set was more accurate. However, I was able to primarily use data sets from Dr. Heilman because they were consistently more complete. The data set from Dr. Sober was then used, when possible, to fill in any missing gaps. Also, graphs comparing the two data sets showed a possible one hour shift between the sets (Figure 3). Later, I was able to contact both Dr. Heilman and Dr. Sober about this issue. I found out that the data from Dr. Heilman was taken at Central Standard Time whereas the data from Dr. Sober was taken at Eastern Standard Time. This accounted for the one hour shift between data from the two researchers. I was also informed that Dr. Sober's data was taken from a height of 1 meter and Dr. Heilman's data was taken from heights of 2 and 10 meters. This was one explanation for the discrepancies between the two data sets, especially for light, temperature, and relative humidity values. For example, a temperature reading on a typical summer day will generally be higher when taken from a height of 10 meters with full sun exposure than from a 1-meter reading with shade cover. Sensor error is another source of inconsistency between the sets, as explained by Dr. Sober. The temperature sensors he uses in separate rings can produce data with up to a half degree Celsius difference. This added to a .3 degrees Celsius conversion error (of temperature units) can result in a total error of up to 1.5 degrees Celsius between temperature sensors.

Next, another obstacle arose when no data was available for a certain time span. To fill in such gaps, data was taken from the same treatment using a different day. However, it was often difficult to find a day that was similar to the one with gaps. This required a search through many sections of data in order to find a day that matched

closely. I would begin by comparing days near to the day with missing data. This was usually sufficient because the experimental variables did not fluctuate greatly from day to day. However, when none of the surrounding days matched the day in question, I continued to look through the entire data file until I found one that corresponded. Due to the randomness of this method, the data filled in this manner is more uncertain.

Other challenges I faced included not knowing how to differentiate between reasonable and unreasonable data. This knowledge would be learned by talking closely with those collecting the data and by working with the FACE Experiment directly. I plan to get more of this type of information this summer through email and a visit to the FACE site. Handling immensely large data files was also a challenge because with such large files it was easy to become confused among specific treatments and replications. However, I was able to meet the challenges with solutions that allowed me to continue my research.

SOFTWARE USED

I received all the data files either in Microsoft Excel or Microsoft Access. I often had to extract specific data for a certain treatment from the files in Microsoft Access. Data that could not be found there was located from the FACE Experiment website (<http://www.nrri.umn.edu/factsii/>). I used Microsoft Excel to fill in and analyze all of the missing data. I also used Microsoft Excel to create the majority of the graphs used for analysis purposes. However, for a number of graphs I was able to use the highly advanced, multi-dimensional graphing program, Tecplot (Figure 4).

CONCLUSIONS

I was fairly effective in accomplishing the educational objectives outlined in my project proposal. Data from the year 1998 was the most incomplete and also the most time consuming to fill as a result. While I completed environmental data files for the years 1998, 1999, and 2000, time did not allow for a detailed statistical analysis of these files. However, I was able to compare and contrast all of the data sets using graphical analysis. Also, an important aspect of my project that I did not anticipate is that my analysis uncovered inconsistencies and issues with various data sets. First, I discovered the possible one hour shift between the data from Dr. Heilman and Dr. Sober. Second, I discovered for certain environmental variables that a number of replications from the experiment contained copied data. Third, I noticed considerable differences between data from Dr. Heilman and Dr. Sober, which I later found to result from the locations and heights at which the data was taken. I have since discovered that it would be very beneficial to have a list of the causes of uncertainties and sources of inaccuracy, as well as a log of lightening strikes and other incidences that occurred that resulted in inaccurate data. I did not have time to begin the second part of my proposed research, analyzing the accuracy of the FACE gas applications recorded in the field data files. I plan to work on the unfinished portions of my project as I continue this line of research for the next several months.

FUTURE WORK

Over the next several months, with additional guidance from Professor Yongcheng Qi, (Statistics, UMD), and Professor Lenz, Professor Stech, and Dr. Host, I plan to use statistical methods for analyzing the raw and the completed FACE data files. The results of analyzing the data files will determine variability of the data values from

data set to data set and whether it is statistically significant. I also plan to find a model or method that will fill in missing data with expected values. Once I am certain that the data files are complete and satisfactorily accurate, I will construct two tables to summarize portions of the data. The first table will show sources and levels of uncertainty and inaccuracy for various data files. The second table will describe incidences such as lightning strikes, power outages, instrumentation failures, or other problems that indicate when and where data from the files should be disregarded. After these tables are complete, I will analyze how the amounts of carbon dioxide and ozone applied to the different treatments compare to the target application values specified in the experimental design. As I continue to work with data files from the FACE Experiment, I believe I will gain new knowledge and experience in working with large data files, databases, and the researchers who collect and record the data. I will also learn to use statistical methods, as well as graphical methods, for analyzing data and communicating the results to others. My findings will be incorporated into a technical report that will be shared with the FACE researchers.

During the week of July 15th through July 19th, I will be attending the biomass harvest at the FACE site in Rhinelander, Wisconsin. At the harvest, I will be involved in taking measurements and weights of various parts of the harvested trees from within the experiment.

EVALUATION

Overall, I learned a great deal from my UROP experience. Not only did I expand my knowledge working with incredibly large data files, but I also gained experience working with various other researchers within and outside of my discipline. I thoroughly

enjoyed being a part of a research group and found hearing about the many different parts of the project being worked on simultaneously very interesting. Altogether, my UROP experience has been very beneficial and I would recommend the Undergraduate Research Opportunities Program to any college student seeking extra knowledge and experience in their field of study.

Par Values 1998, Daily Average

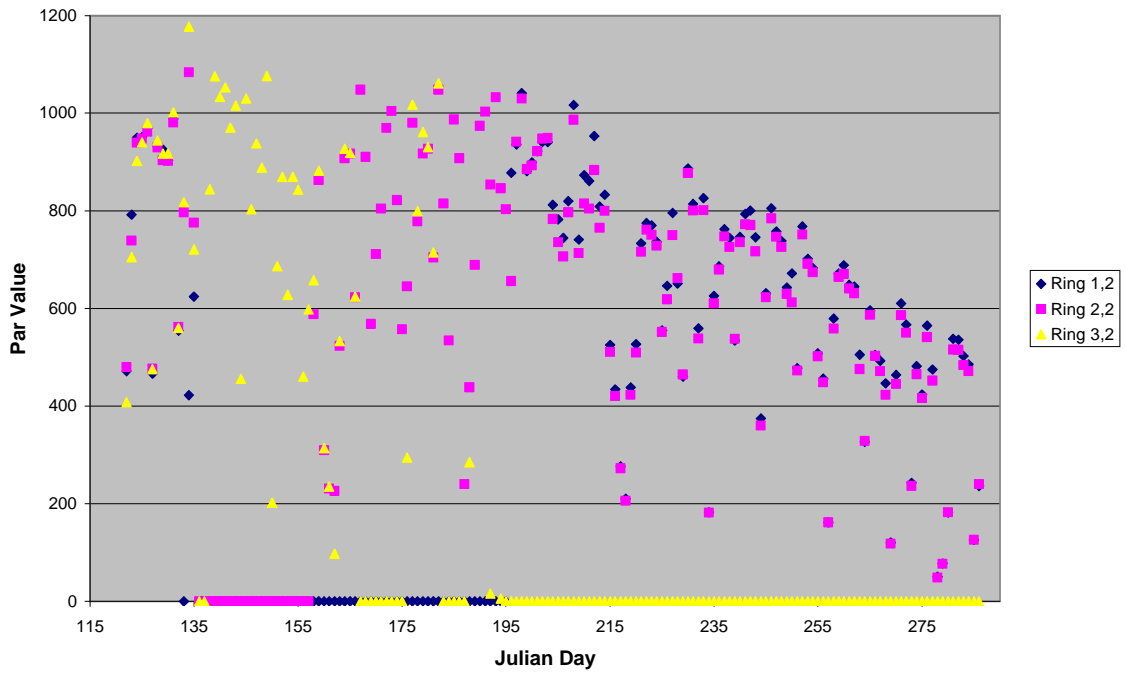


Figure 1: Light values (PAR) for three different treatments in 1998, missing values shown by data at bottom of graph.

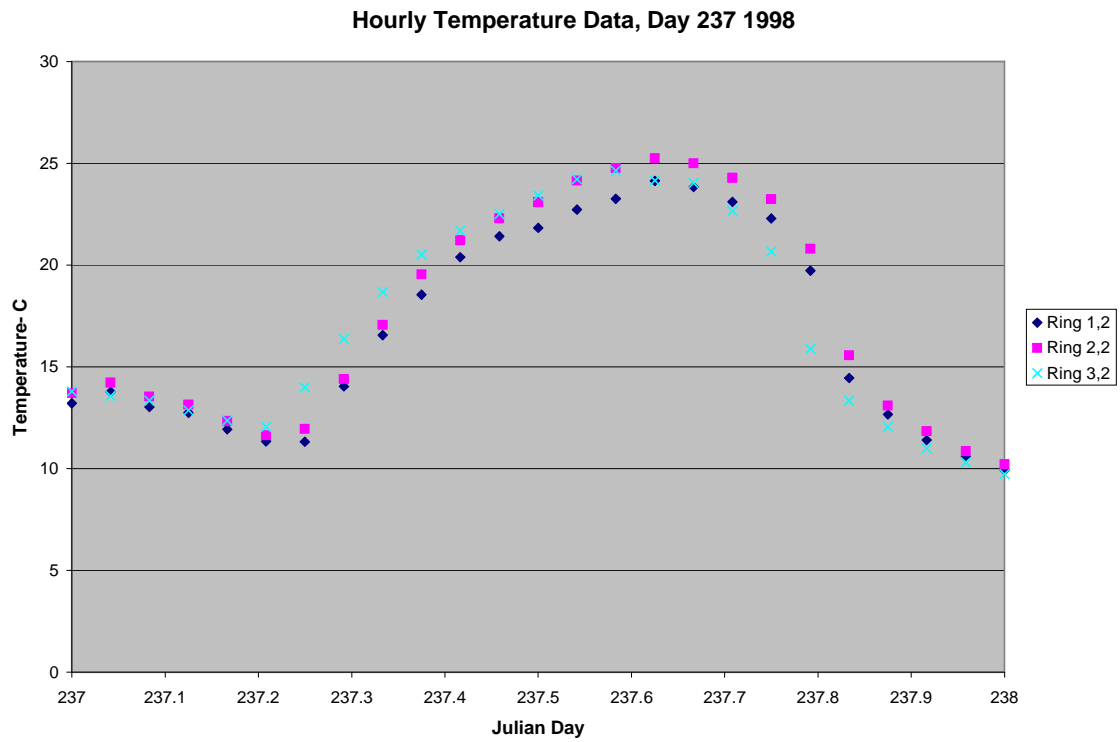


Figure 2: Temperature values for three different treatments in 1998, showing a graphical comparison used to find a close match.

Hourly Par Values, Ring 3,4 Day 126-1998

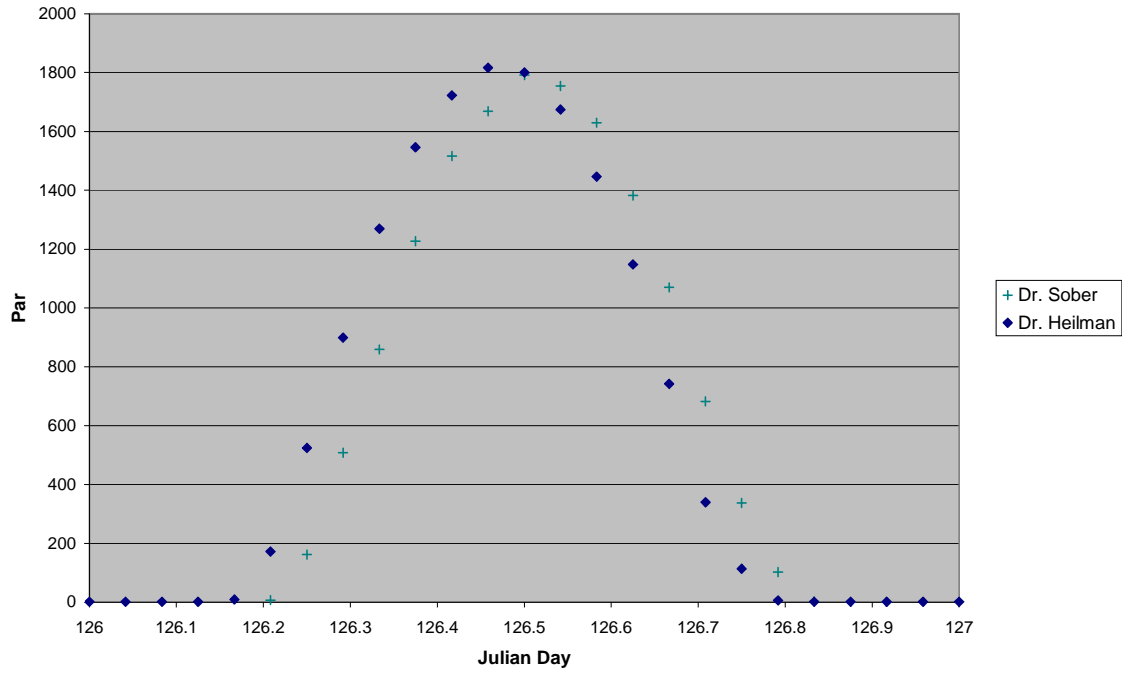


Figure 3: Light values from two different researchers in 1998, Dr. Sober and Dr. Heilman, showing a one-hour shift between data sets.

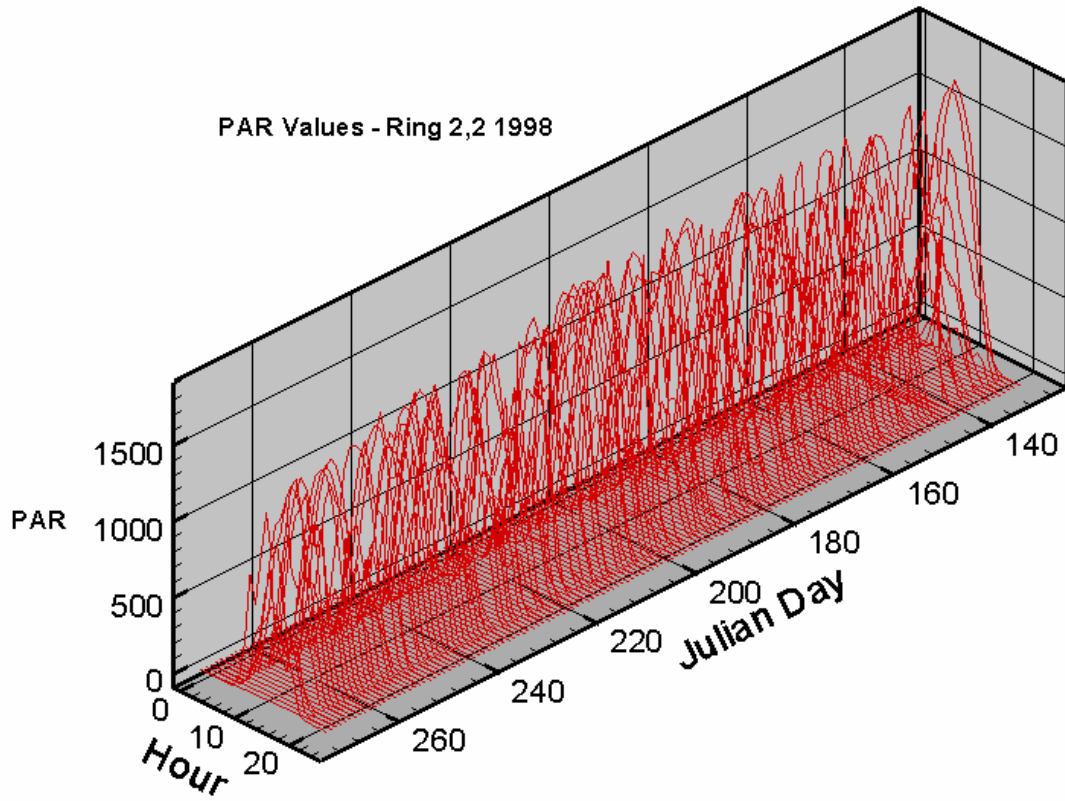


Figure 4: Light values from Ring 2,2 in 1998, showing the capabilities of a three-dimensional graph in Tecplot.

Dickson, R.E., Lewin, K.F., Isebrands, J.G., Coleman, M.D., Heilman, W.E., Riemenschneider, D.E., Sober, J., Host, G.E., Zak, D.R., Hendrey, G.R., Pregitzer, K.S. and Karnosky, D.F. (2000). *Forest Atmosphere Carbon Transfer and Storage (FACTS-II) The Aspen Free-air CO₂ and O₃ Enrichment (FACE) Project: An Overview*.

FACTS II: The Aspen FACE Experiment. Retrieved from <http://www.nrri.umn.edu/factsii/>.

Lenz, K.E., H.W. Stech. (2000). Student Research Participation in Multidisciplinary Tree-Soil-Atmosphere Modeling. *Proceedings of the 2000 Conference on Mathematical Modeling in the Undergraduate Curriculum, University of Wisconsin-La Crosse*.

Roskoski, Kyle. (2002). Converting Weather Data to an Hourly Format Using Queries and Macros. In progress.