Measurement, Uncertainty, and Data Analysis
Activity #2: Quantifying the Effect of Large Data Sets
Student Worksheet

Objective

The purpose of this experiment is to use MOSAIC data to determine the effect of averaging many measurements on the uncertainty associated with those measurements.

Background

The MOSAIC system consists of a Direct TV-style radio dish directed 8° above Earth's horizon combined with a receiver that acts as a spectrometer. Because the receiver is sensitive to a frequency corresponding to a rotational transition of ozone, it is able make measurements of the spectral signal coming from ozone located in the mesosphere.

Mesospheric ozone is destroyed every day under the ultraviolet light of the sun and created again every night when the sunlight goes away, so the signal is strongest during the nighttime hours. Also, some seasonal variations have been observed, with highest ozone concentrations in the spring and fall and lowest concentrations in the summer and winter. Five locations (CHS, Ridge, Bridgewater, Union, UNCG) host a MOSAIC system (since 2009, Day 171), but all point at the same place in the sky. A sixth location (Lynnfield High School) has been collecting data on the same place in the sky since August 6, 2010, or 2010, Day 218). This means that it is reasonable to average the signal across different spectrometers for the same day, but when averaging is done over long time periods (comparable to a season or about 90 days), variations in actual ozone intensity may be detected, resulting in changes from day to day on the actual level of ozone.

The web interface to access the data collected by the MOSAIC systems across the east coast of the US is available at http://www.haystack.mit.edu/ozone/. This activity will require students to create ozone spectrum plots corresponding to different lengths of time, and study the effect of increasing the sample length on the error associated with the peak value.

Method

Go to the MOSAIC user interface (http://www.haystack.mit.edu/ozone). You will be plotting ozone spectra for different numbers of days and numbers of spectrometers within a time interval assigned by your teacher. Select the box labeled "plot nighttime data only" so that you get the strongest signal associated with each day. Assigned time interval: ________________________

Use the web interface to produce ozone spectra for a wide range of number of days and number of spectrometers within the time interval. Collect data from at least 10-15 combinations. Select the box for "nighttime data only" in order to get a stronger signal (remember, ozone is destroyed in the presence of sunlight in the mesosphere every day). After producing each spectrum, click on the "sum.txt" file to the right of the graph in order to record the parameters of the curve fit to the data. Specifically, you should record the "peak" value and the "rms resid" value.

Enter your data in a table that looks like this:

# days	**# sites**	**# "data days"**	**peak**	**rms resid**

You can have Excel compute the “# data days” column for you by simply multiplying the number of days by the number of sites.

Data Analysis

Answer the following questions in complete sentences as part of your lab report. Turn them in, along with your graph and data table.

1. Make a graph of your data. What is the independent variable? What is the dependent variable?
2. Once you have made a scatter graph, label the axes of your graph and give it a title.
3. What trend does the data exhibit? Describe it with as much specificity as you can.
4. Fit a trendline to your data, and display the equation and R^2 value. Change x and y to appropriate variables to reflect the actual quantities being graphed.

Discussion/Results

Write a paragraph summarizing your findings and stating your results. Be sure to address the following in your paragraph (but don’t number your answers!).

1. Refer to the objective of this experiment, and state your results.
2. How did the appearance of the spectra change as you added more sites and more days to the analysis?
3. How does uncertainty depend on the number of measurements being averaged together?
4. What are the limitations of averaging together data from multiple sites and multiple days? Why were we able to do it in this case? Why might it not be a good idea to average over a whole year?
5. Did you verify the expected relationship between standard error and number of measurements? (If you aren’t sure what the expected relationship is, check your notes!) If not, why not?

Measurement, Uncertainty, and Data Analysis
Activity #2: Quantifying the Effect of Large Numbers
Teacher Notes

Teacher Information

Background

The MOSAIC system consists of a Direct TV-style radio dish directed 8° above Earth's horizon combined with a receiver that acts as a spectrometer. Because the receiver is sensitive to a frequency corresponding to a rotational transition of ozone, it is able make measurements of the spectral signal coming from ozone located in the mesosphere.

Mesospheric ozone is destroyed every day under the ultraviolet light of the sun and created again every night when the sunlight goes away, so the signal is strongest during the nighttime hours. Also, some seasonal variations have been observed, with highest ozone concentrations in the spring and fall and lowest concentrations in the summer and winter. Five locations (CHS, Ridge, Bridgewater, Union, UNCG) host a MOSAIC system (since 2009, Day 171), but all point at the same place in the sky. A sixth location (Lynnfield High School) has been collecting data on the same place in the sky since August 6, 2010, or 2010, Day 218). This means that it is reasonable to average the signal across different spectrometers for the same day, but when averaging is done over long time periods (comparable to a season or about 90 days), variations in actual ozone intensity may be detected, resulting in changes from day to day on the actual level of ozone.

The web interface to access the data collected by the MOSAIC systems across the east coast of the US is available at http://www.haystack.mit.edu/ozone/. This activity will require students to create ozone spectrum plots corresponding to different lengths of time, and study the effect of increasing the sample length on the error associated with the peak value.

Purpose

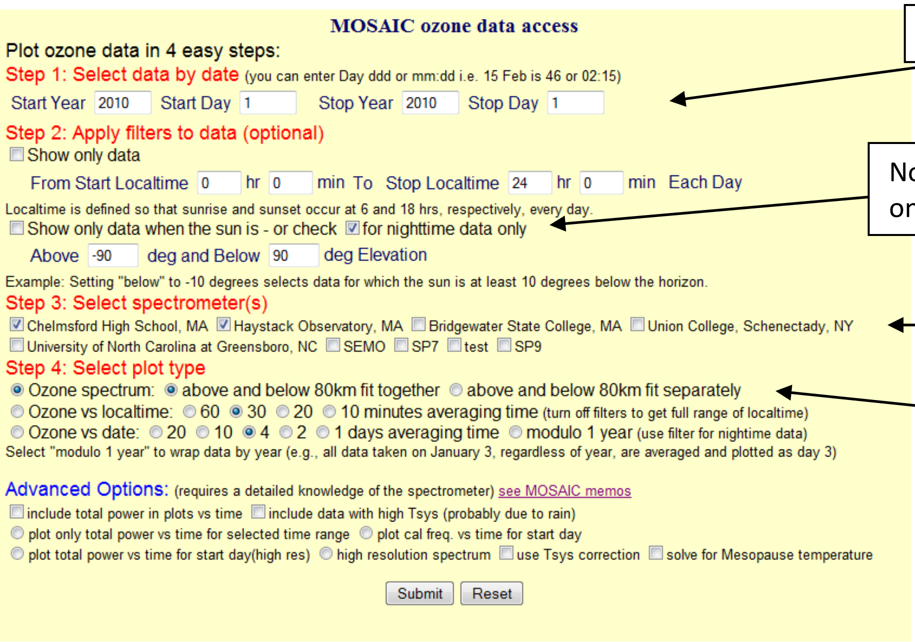
By completing this lab, students will discover the advantage of averaging together signals from multiple sources. They should be able to verify the quantitative relationship between number of samples and the standard error associated with a measurement. They should learn a bit about the limitations of averaging multiple sources when the signal itself is changing.

Set Up and Preparation

Students will need access to computers with internet access, along with Excel (or some other graphical program). After students have been introduced to the basics of the MOSAIC system and the basics of uncertainty and distributions (from the Power Point "Data Analysis" or other materials), they will be ready to tackle this activity. Students should also be familiar with residuals, and how they relate to fitting data to lines or curves.

Data Collection

Using the Web Interface, have students collect data for different lengths of time and from different numbers of spectrometers. This will serve as a measure of "sample size" within the distribution, and will be the independent variable for the experiment. The dependent variable will be the rms of the residual to the plot fit, which can be thought of as simply a measure of the uncertainty associated with the spectra.
This is the input screen for 1 day and 2 sites (2 "data days"):



MOSAIC ozone data access

Plot ozone data in 4 easy steps:

Step 1: Select data by date (you can enter Day ddd or mm:dd i.e. 15 Feb is 46 or 02:15)

Start Year 2010 Start Day 1 Stop Year 2010 Stop Day 1

Step 2: Apply filters to data (optional)

□ Show only data

From Start Localtime 0 hr 0 min To Stop Localtime 24 hr 0 min Each Day

Localtime is defined so that sunrise and sunset occur at 6 and 18 hrs, respectively, every day.

□ Show only data when the sun is - or check ☑ for nighttime data only

Above -90 deg and Below 90 deg Elevation

Example: Setting "below" to -10 degrees selects data for which the sun is at least 10 degrees below the horizon.

Step 3: Select spectrometer(s)

☑ Chelmsford High School, MA ☑ Haystack Observatory, MA □ Bridgewater State College, MA □ Union College, Schenectady, NY
□ University of North Carolina at Greensboro, NC □ SEMO □ SP7 □ test □ SP9

Step 4: Select plot type

● Ozone spectrum: ● above and below 80km fit together ○ above and below 80km fit separately
○ Ozone vs localtime: ○ 60 ● 30 ○ 20 ○ 10 minutes averaging time (turn off filters to get full range of localtime)
○ Ozone vs date: ○ 20 ○ 10 ● 4 ○ 2 ○ 1 days averaging time ○ modulo 1 year (use filter for nightime data)
Select "modulo 1 year" to wrap data by year (e.g., all data taken on January 3, regardless of year, are averaged and plotted as day 3)

Advanced Options: (requires a detailed knowledge of the spectrometer) see MOSAIC memos

□ include total power in plots vs time □ include data with high Tsys (probably due to rain)
○ plot only total power vs time for selected time range ○ plot cal freq. vs time for start day
○ plot total power vs time for start day(high res) ○ high resolution spectrum □ use Tsys correction □ solve for Mesopause temperature

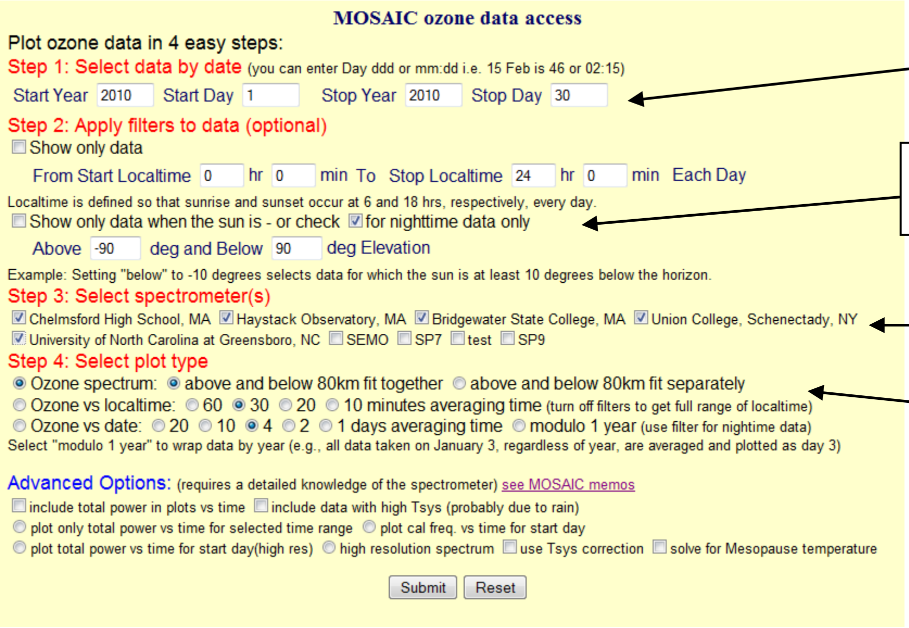
1 Day Selected (2010, Day 1)

Note filter for "nighttime data only" is selected.

2 sites selected (Chelmsford and Haystack)

Ozone spectrum is selected, with above and below 80 km fit together.

This is the input screen for 30 days and 5 sites (150 "data days")



MOSAIC ozone data access

Plot ozone data in 4 easy steps:

Step 1: Select data by date (you can enter Day ddd or mm:dd i.e. 15 Feb is 46 or 02:15)

Start Year 2010 Start Day 1 Stop Year 2010 Stop Day 30

Step 2: Apply filters to data (optional)

□ Show only data

From Start Localtime 0 hr 0 min To Stop Localtime 24 hr 0 min Each Day

Localtime is defined so that sunrise and sunset occur at 6 and 18 hrs, respectively, every day.

□ Show only data when the sun is - or check ☑ for nighttime data only

Above -90 deg and Below 90 deg Elevation

Example: Setting "below" to -10 degrees selects data for which the sun is at least 10 degrees below the horizon.

Step 3: Select spectrometer(s)

☑ Chelmsford High School, MA ☑ Haystack Observatory, MA ☑ Bridgewater State College, MA ☑ Union College, Schenectady, NY
☑ University of North Carolina at Greensboro, NC □ SEMO □ SP7 □ test □ SP9

Step 4: Select plot type

● Ozone spectrum: ● above and below 80km fit together ○ above and below 80km fit separately
○ Ozone vs localtime: ○ 60 ● 30 ○ 20 ○ 10 minutes averaging time (turn off filters to get full range of localtime)
○ Ozone vs date: ○ 20 ○ 10 ● 4 ○ 2 ○ 1 days averaging time ○ modulo 1 year (use filter for nightime data)
Select "modulo 1 year" to wrap data by year (e.g., all data taken on January 3, regardless of year, are averaged and plotted as day 3)

Advanced Options: (requires a detailed knowledge of the spectrometer) see MOSAIC memos

□ include total power in plots vs time □ include data with high Tsys (probably due to rain)
○ plot only total power vs time for selected time range ○ plot cal freq. vs time for start day
○ plot total power vs time for start day(high res) ○ high resolution spectrum □ use Tsys correction □ solve for Mesopause temperature

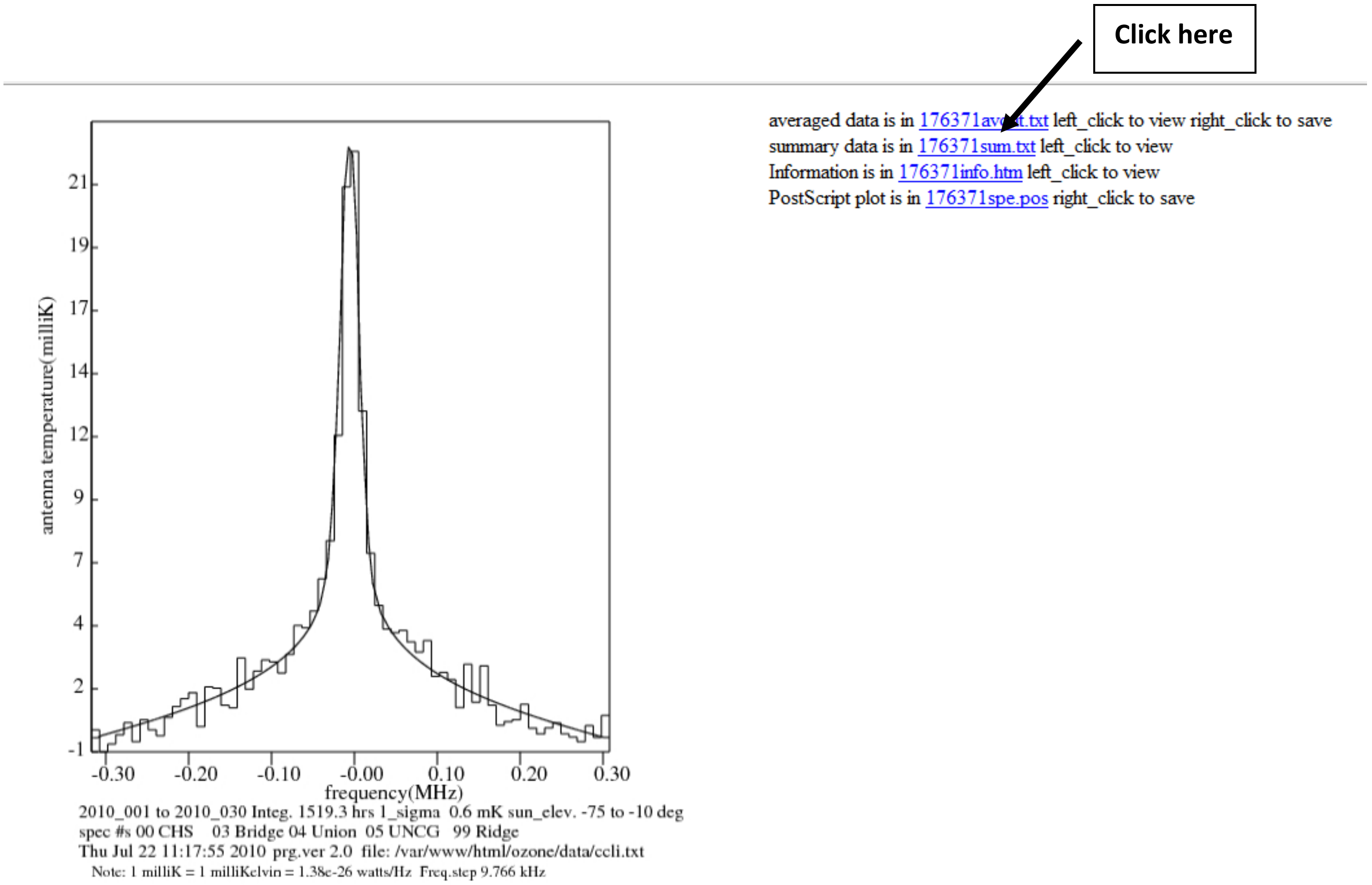
30 Days Selected (2010, Day 1-30)

Note filter for "nighttime data only" is selected.

5 sites selected (Chelmsford, Haystack, Bridgewater, Union, UNCG)

Ozone spectrum is selected, with above and below 80 km fit together.

It shouldn't matter which spectrometers they choose, but they should vary (and record) the number they are using.) Once they produce a spectrum, click on the sum.txt file, as shown below.

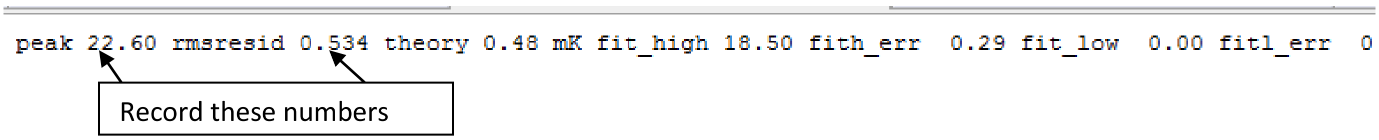


Click here

averaged data is in 176371avgdt.txt left_click to view right_click to save
summary data is in 176371sum.txt left_click to view
Information is in 176371info.htm left_click to view
PostScript plot is in 176371spe.pos right_click to save

2010_001 to 2010_030 Integ. 1519.3 hrs 1_sigma 0.6 mK sun_elev. -75 to -10 deg
spec #s 00 CHS 03 Bridge 04 Union 05 UNCG 99 Ridge
Thu Jul 22 11:17:55 2010 prg.ver 2.0 file: /var/www/html/ozone/data/ccli.txt
Note: 1 milliK = 1 milliKelvin = 1.38e-26 watts/Hz Freq.step 9.766 kHz

The sum.txt file will look like what is shown below (but it has been cut off, since the data at the end of the line is not important for this experiment). It provides the parameters of the curve fit to the data. Have students record the "peak" and "rmsresid" value for each trial.



```
peak 22.60 rmsresid 0.534 theory 0.48 mK fit_high 18.50 fith_err  0.29 fit_low  0.00 fitl_err  0
```

Record these numbers

Have students collect 10-15 data points, covering a wide range of values for number of days and number of sites. They should have one data point that is all 5 sites and about 40-50 days, and another that corresponds to a single site on a single day.

You may want to assign different groups different time intervals to choose their data from. Just make sure that each time interval is one through which all 5 sites were operational (or let students know that they have fewer than five sites to choose from).

Some possible time intervals to assign to student groups:
2009, Day 180- 230
2009, Day 285 – 335
2009, Day 335 – 2010, Day 20
2010, Day 20 – 70
2010, Day 70 – 120

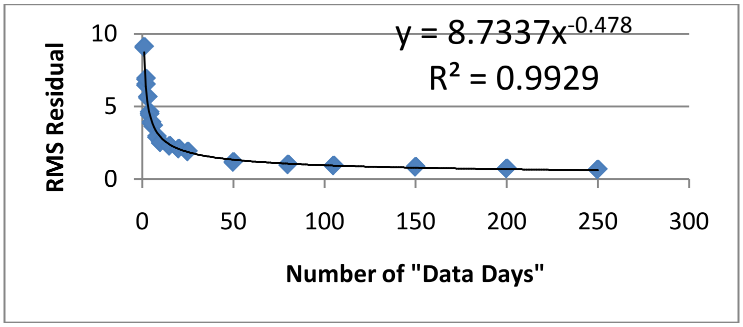
[It is probably best to avoid the time between Days 235- 285 in 2009 and Days 120 – 170 in 2010 since the ozone is changing very quickly in these intervals]

Students will produce a data table that looks something like this:

# days	# sites	peak	rms resid
2	1	11.7	6.505
2	2	11.21	4.603
2	3	16.61	3.676
2	4	17.44	2.933
2	5	17.48	2.54
3	5	17.63	2.297
4	5	19.28	2.107
5	5	18.42	1.909
1	5	21.13	3.905
1	1	21.11	9.101
1	2	18.32	6.915
1	3	23.38	5.647
1	4	20.99	4.466
1	5	21.13	3.905
21	5	21.91	0.948
10	5	20.47	1.189
50	5	19.35	0.672
30	5	21.48	0.857
40	5	20.28	0.744
40	2	18.86	1.02

Then, have students compute (with the aid of Excel, perhaps) the total number of "data days" for each spectrum by multiplying the number of days by the number of sites.

Have students graph "data days" on the x-axis and the rms residual on the y-axis. You should see a nice, smooth, decreasing function. If you fit a power and display the equation, you should find it very closely resembles the 1/√N relationship expected for large sets of data, as below.



Note: It is likely students will find that the power on their trendline is slightly greater than -0.500 (that is, between -0.400 and -0.500). This is because the uncertainty in the measurements depends a bit less strongly on the number of trials than you would expect due only to the number of trials being averaged.