Initial analysis of the distribution of selected species at Seney.

Peter Scheff
Environmental and Occupational Health Science
University of Illinois at Chicago

Following is an initial look at the distribution of selected species monitored at the Seney air monitoring station. Seney was selected for this initial analysis because it has the greatest number of samples and highest sampling frequency. In the data set analyzed for Seney, there were approximately 169 samples collected from November 1, 2001 through July 6, 2002 at a rate of 4 samples every 6 days.

The plots are organized in groups of three. Each group looks at a distribution, or group of distributions based on 1) all samples (4 samples in 6 days), 2) two samples in 6 days, and 3) one sample in 6 days. The approach is to compare these three distributions to visually evaluate the effect of sampling frequency on observed concentrations. If the underlying distribution is correctly reproduced at lower sampling frequencies, then sampling at the high frequency may not be necessary.

Figures 1, 2 and 3 show box plots for major components (ammonium, EC, OC, NO3, SO4, and PM2.5) from the speciation monitoring program at Seney. As stated above, figure 1 contains approximately 169 observation collected on 4 days during each 6 day cycle. Figure 2 contains approximately 84 observation (samples collected on day 1 and day 4 of each 6 day cycle) and Figure 3 contains approximately 42 samples (collected on day 1 of each 6 day cycle). These figures show that box and whiskers are relatively unchanged by reducing the sampling frequency.

To allow for improved comparison at low concentrations, these figures are re plotted in Figures 4, 5, and 6 on a log-scale. These figures show that the distributions are unchanged when sampling frequency is reduced. Assuming that Figure 4 best represents the full underlying distributions of concentrations (one sample every day), dropping down to 1 sample every 3 days (Figure 5) and 1 sample every 6 days (Figure 6) shows that the underlying distribution is correctly estimated at the lower sampling frequencies.

These first 6 figures show that the only difference due to sampling frequency is that the extreme values are better represented at higher sampling frequencies. If the primary objective of the monitoring network is to capture a representative distribution of concentrations, then sampling 1 in 3 days, or 1 in 6 days does an excellent job. However, if the primary objective is to capture the extreme value, then one is always more likely to do that at a high sampling frequency.

The next three figures look at the actual distribution of PM2.5. These plots are of the expected value assuming a normal distribution versus the concentration on a log-scale. If data on this plot is close to a straight line, then one can conclude that the observations are log-normally distributed. For PM2.5 at the three sampling frequencies (Figures 7, 8, and 9), the data are very close to the log-normal distribution. The least-squares fit line on these figures represents a fit of the data to the log-normal distribution. The value of x for an expected value of 0 is the geometric mean, and the slope of the line is the geometric standard deviation. As shown on these three figures, the lines are almost identical. This demonstrates

that a reduced sampling frequency does not affect the measurement of the distribution of PM2.5 values at Seney.

Figures 10, 11 and 12 repeat the analysis of distribution for nitrate. While the data are not as consistent with the log-normal distribution as was PM2.5, the slopes and intercepts of the three lines are almost exactly the same. This also demonstrates that a reduced sampling frequency does not affect measured distribution of nitrate at Seney. Figures 13, 14 and 15 repeat this analysis for organic carbon and support the same conclusion.

Finally, figures 13, 14 and 15 are box plots of selected elements at the three sampling frequencies. As was the case for figures 4, 5 and 6, except for extreme values, the three plots are very similar.

Taken as a whole, this analysis shows that a one in 6 day sampling frequency is able to accurately represent the distribution of the concentrations of selected species at Seney. From this analysis, it will be possible to fit the observations to a log-normal distribution and correctly estimate the geometric mean and standard deviation. Measuring all of the extreme values is not necessary. However, as sampling frequency is reduced, the likelihood of actually measuring the extreme values is reduced.

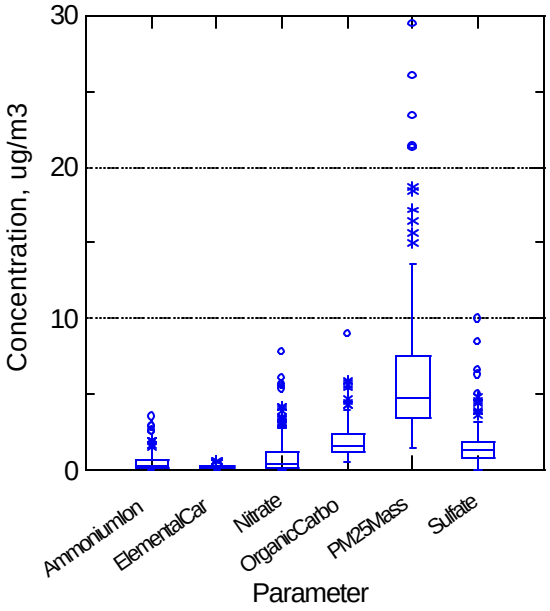

Figure 1 All samples (4 in 6 days), Seney

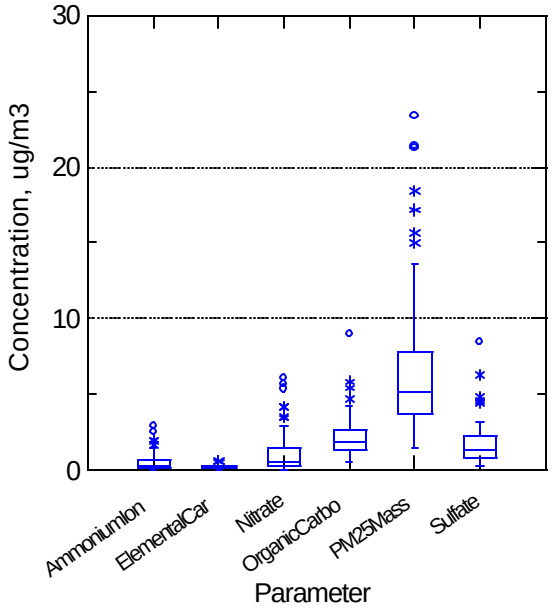

Figure 2 Two samples in 6 days, Seney

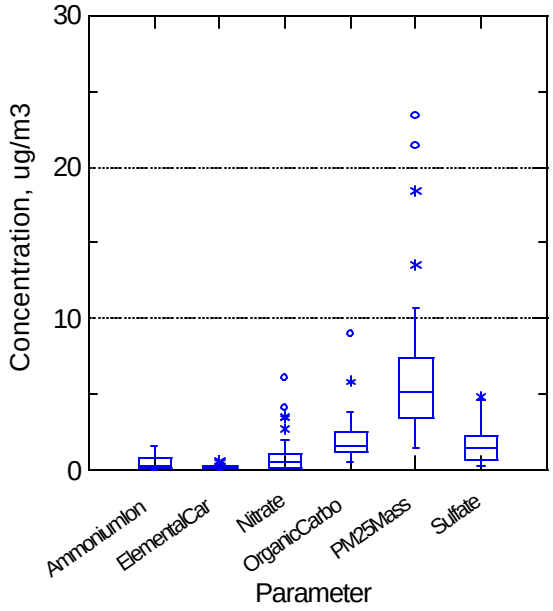

Figure 3 One sample in 6 days, Seney

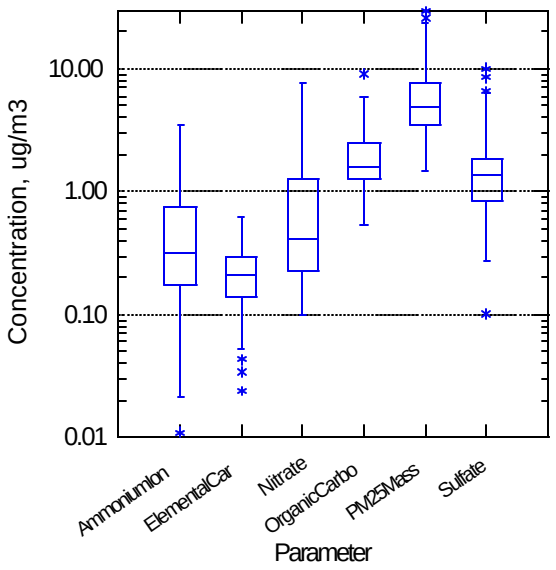

Figure 4 All samples (4 in 6 days), Seney

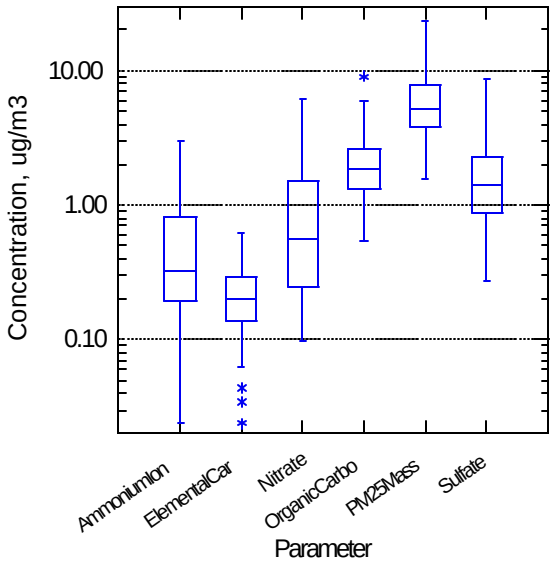

Figure 5 Two samples in 6 days, Seney

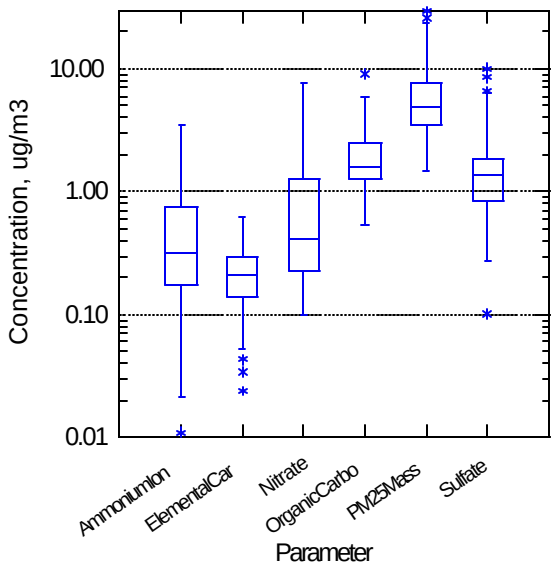

Figure 6 One sample in 6 days, Seney

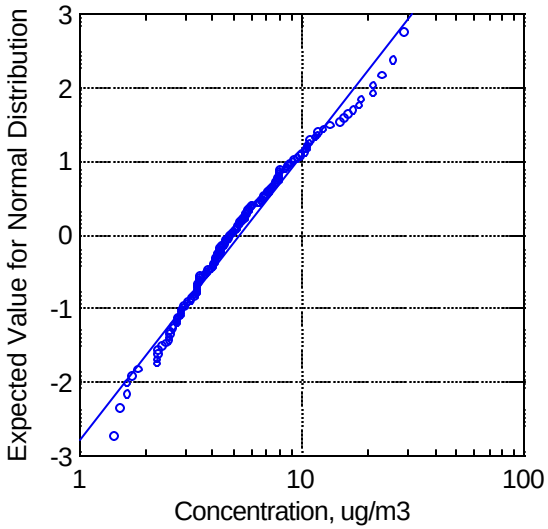

Figure 7 All samples, PM2.5, Seney

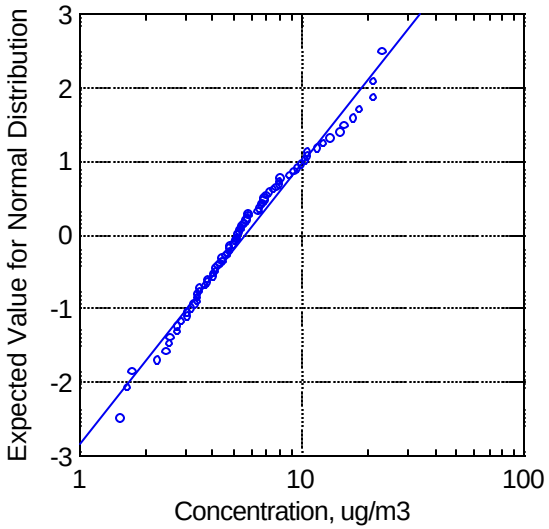

Figure 8 2 samples in 6 days, PM2.5, Seney

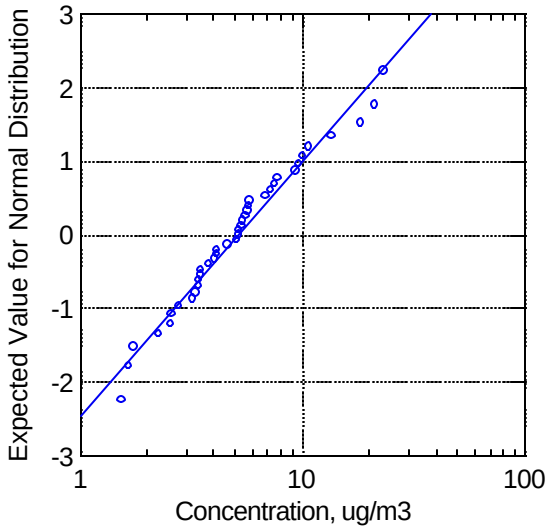

Figure 9 1 sample in 6 days, PM2.5, Seney

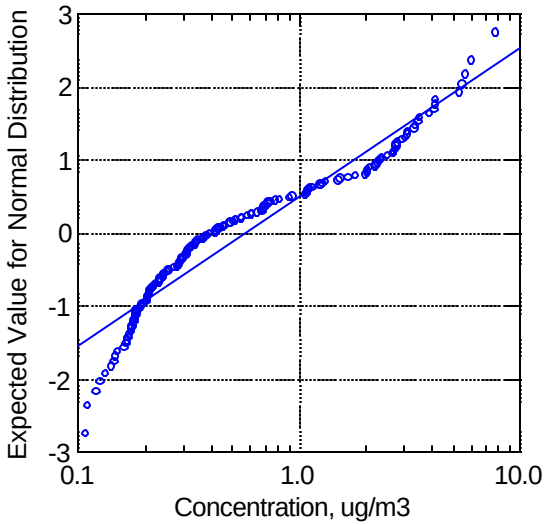

Figure 10 All samples, nitrate, Seney

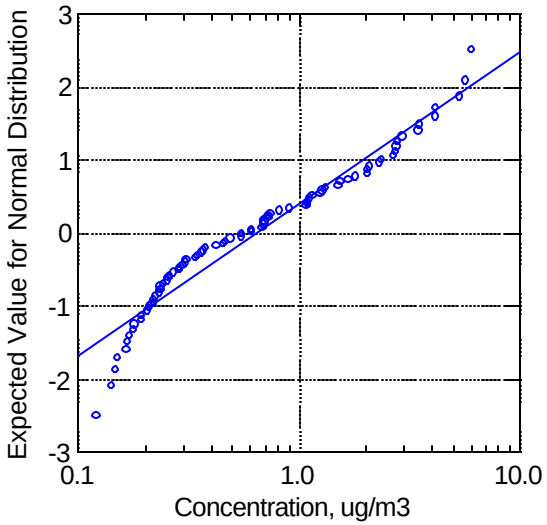

Figure 11 2 samples in 6 days, NO3, Seney

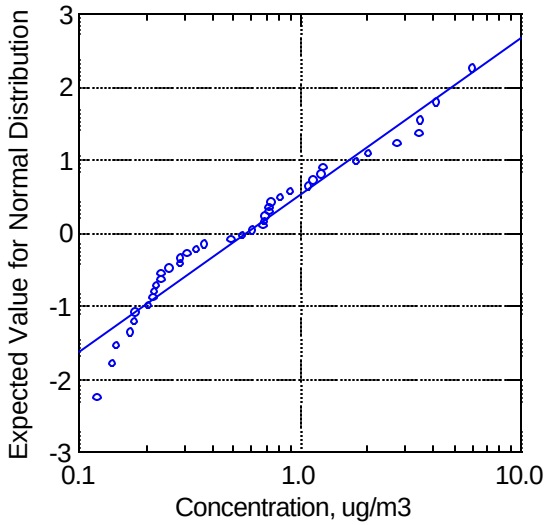

Figure 12 1 sample in 6 days, NO3, Seney

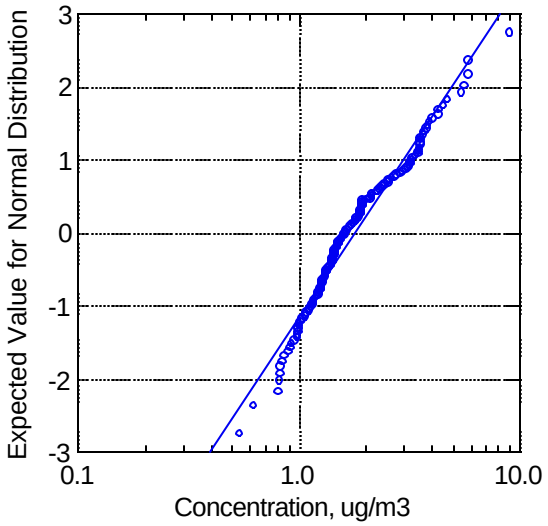

Figure 13 All samples, OC, Seney

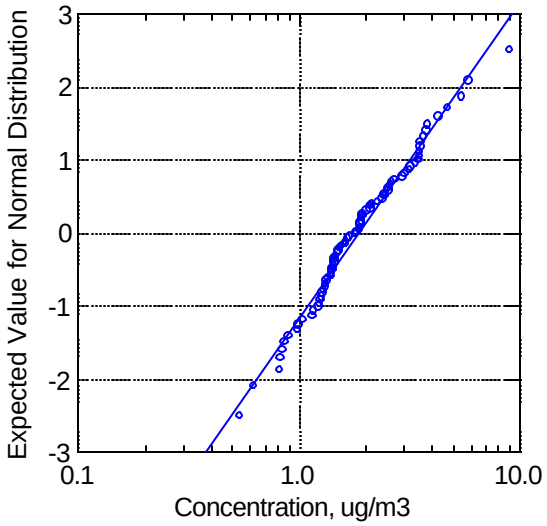

Figure 14 2 samples in 6 days, OC, Seney

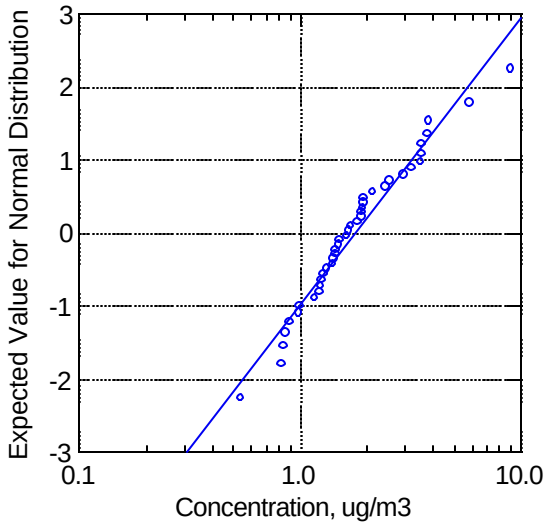

Figure 15 1 sample in 6 days, OC, Seney

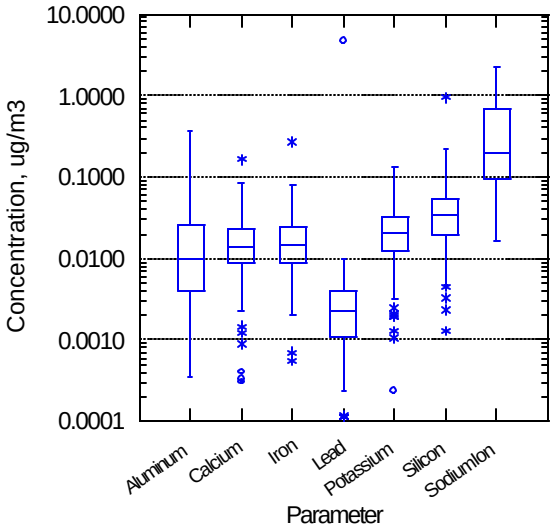

Figure 16 All samples, (4 in 6 days) Seney

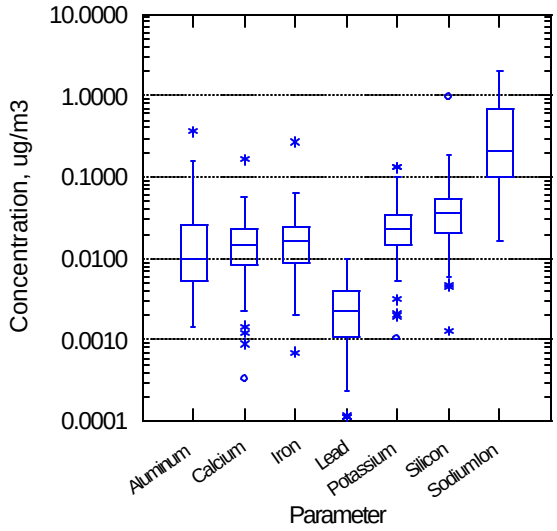

Figure 17 Two samples in 6 days, Seney

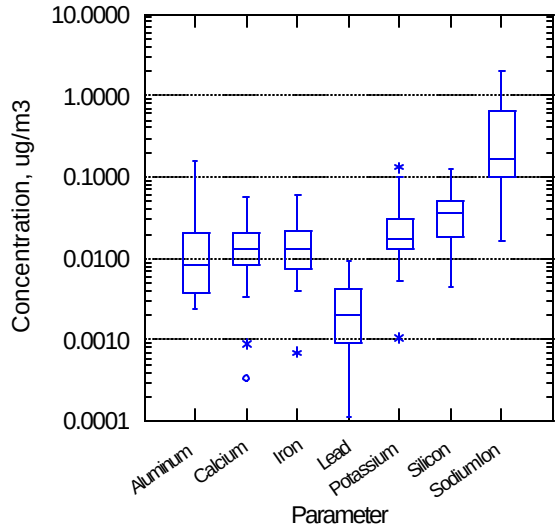

Figure 18 One sample in 6 days, Seney