APPENDIX J
Graphical Tools

J-1. Introduction. Graphs are powerful data evaluation tools, providing a quick assessment of concentration ranges, extreme concentrations or data anomalies, and patterns and trends that may be unapparent otherwise. In exploratory data analysis, various graphical techniques are used initially to display the data so that users may determine what statistical evaluations will be used. Although a subjective assessment of a plot alone is often inadequate to make conclusions about the significance of a trend or association, plots support quantitative statistical tests.

J-1.1. This Appendix presents some common graphical methods for presenting environmental data in meaningful ways. These graphical methods are:

- Histogram/Frequency Plots.
- Box-and-Whiskers Plots.
- Quantile Plots.
- Normal Probability Plots (Quantile-Quantile Plots).
- Empirical Quantile-Quantile Plots.
- Plots for Temporal Data.
- Plots for Spatial Data.
- Plots for Two or More Variables.
- Contouring Data.

J-1.2. Additional information on most of the plots presented here may be found in Mason et al. (1989). For temporal and spatial plots see EPA 600/R-96/084, QA/G-9.

J-2. Histogram/Frequency Plots

J-2.1. *Introduction*. Two of the oldest methods for summarizing data distributions are the frequency plot (Figure J-1) and histogram (Figure J-2). Both frequency plots and histograms divide the range of measured values of a variable into equal intervals, and use a bar graph to display the results. In a frequency plot, the height of each bar represents the number of observations within each interval. In a histogram, the height of each bar represents the percentage of observations within each interval.

J-2.1.1. There are slight differences between the histogram and the frequency plot. In the frequency plot, the *relative height of the bars* represents the relative density of the data or number of observations within a group. In a histogram, the *area within the bar* represents the relative density of the data or percentage of observations within a group.

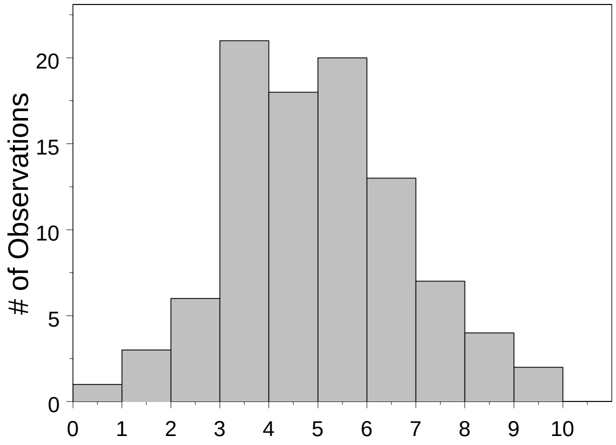


Figure J-1. Frequency plot: normal data.

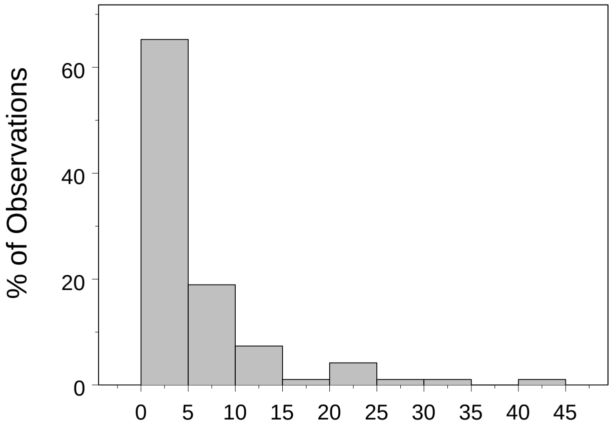


Figure J-2. Histogram: lognormal data.

J-2.1.2. When plotting a histogram for a continuous variable (such as concentration), it is necessary to decide on an endpoint convention, that is, what to do with cases that fall on the boundary of a box. With discrete variables (i.e., family size), the intervals can be centered in between the variables. For the family size data, the intervals can span between 1.5 and 2.5, 2.5 and 3.5, and so on, so that the whole numbers that relate to the family size can be centered within the box. *The visual impression conveyed by a histogram or a frequency plot can be quite sensitive to the choice of interval width.* The choice of the number of intervals determines whether the histogram shows more detail for small sections of the data or whether the data will be displayed more simply as a smooth overview of the distribution. For a continuous measurement variable, *X*, the histogram should approach the "true" probability distribution as the sample size increases and the width of the intervals decrease. For example, if the variable *X* is normally distributed, then the histogram will approach a Gaussian curve (see Appendix F). Figure J-1 plots 95 observations from a sample from a normal distribution with a mean of 5 and a standard deviation of 2. Notice how the histogram approximates a normal curve. Likewise, Figure J-2 plots 95 observations from a sample from a lognormal distribution with $\mu = 1$ and $\sigma = 1$.

J-2.1.3. Directions for generating a histogram and a frequency plot are presented in Paragraph J-2.2 and an example is contained in Paragraph J-2.3.

J-2.2. *Directions for Generating a Histogram and a Frequency Plot.* Let $x_1, x_2,..., x_n$ represent the *n* data points. To develop a histogram or a frequency plot do the following.

J-2.2.1. Select intervals that cover the range of observations. If possible, these intervals should have equal widths. A rule of thumb is to have between 7 to 11 intervals. If necessary, specify an endpoint convention, i.e., what to do with cases that fall on interval endpoints.

J-2.2.2. Compute the number of observations within each interval. For a frequency plot with equal interval sizes, the number of observations represents the height of the boxes on the frequency plot.

J-2.2.3. Determine the horizontal axis based on the range of the data. The vertical axis for a frequency plot is the number of observations. The vertical axis of the histogram is the percentage (or proportion) of results that fall within each interval on the x-axis.

J-2.2.4. For a histogram, compute the percentage of observations within each interval by dividing the number of observations within each interval (Step J-2.2.3) by the total number of observations.

J-2.2.5. For a histogram, select a common unit that corresponds to the *x*-axis (Step J-2.2.1). Compute the number of common units in each interval and divide the percentage of observations within each interval (Step J-2.2.4) by this number. This step is only necessary when the intervals (Step J-2.2.1) are not of equal widths.

J-2.2.6. Using boxes, plot the intervals against the results of Step J-2.2.5 for a histogram or the intervals against the number of observations in an interval (Step J-2.2.2) for a frequency plot.

J-2.3. *Example of a Histogram and a Frequency Plot*. Consider the following results of benzene concentrations in groundwater (ppb): 0.0292, 0.0300, 0.0300, 0.0300, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0444, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0469, 0.0469, 0.0469, 0.0480, 0.0504, 0.0504, 0.0504, 0.0548, 0.0585, 0.0605, 0.0605, 0.0605, 0.0641, 0.0641, 0.0641, 0.0641, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0786, 0.0786, 0.0855, 0.0970, 0.0971, 0.1430, 0.2220, 0.2350, 0.3080, 0.4840, 0.6350, 0.7590, 0.8130, 1.1500, 1.7200, 1.7800, 1.8400, 1.8500, 1.9200, 2.0000, 2.0100, 2.1700, 2.1900, 2.3100, 2.4600, 2.6800, 2.7500, 2.9500, 3.4200, 3.4500, 3.7900, 4.3000, 5.4700, 5.7700, 5.8700, 6.1700, 6.9100, 7.2400, 7.5600, 8.3400, 8.6400, 9.3300, 11.000, 11.100, 12.200, 14.100, 17.000, 20.200, 21.800, 29.100, 36.700 and 44.500.

J-2.3.1. These data values span 0 to 50 ppb. Equally sized intervals of 5 ppb will be used: 0 to 5 ppb, 5 to 10 ppb, etc. The endpoint convention will be that values are placed in the highest interval containing the value. For example, a value of 5 ppb will be placed in the interval 5 to 10 ppb instead of 0 to 5 ppb. Table J-1 shows the number of observations within each interval defined here

Table J-1.
Number of Observations within Each Interval

Interval (ppb)	Observations in Interval	Percent Observations in Interval	Percent Observations per ppb
0 5	88	81.5	16.3
5–10	10	9.26	1.85
10–15	4	3.70	0.74
15–20	1	0.926	0.185
20–25	2	1.85	0.370
25–30	1	0.926	0.185
30–35	0	0	0
35–40	1	0.926	0.185
40–45	1	0.926	0.185

J-2.3.2. The horizontal axis for the data is from 0 to 50. The vertical axis for the frequency plot is from 0 to 88 and the vertical axis for the histogram is from 0 to 81.5%.

J-2.3.3. There are 108 observations total, so the number of observations shown in the table will be divided by 108. The results are shown in the third column of the table.

J-2.3.4. A common unit for this data is 1 ppb. In each interval there are five common units so the percentage of observations (third column of the table) should be divided by 5 (fourth column).

J-2.3.5. The frequency plot (Figure J-3) and the histogram (Figure J-4) are shown below.

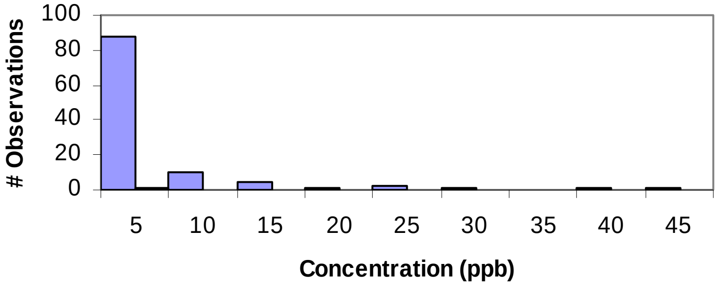


Figure J-3. Frequency plot.

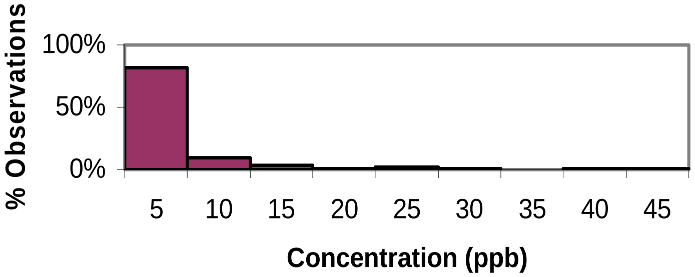


Figure J-4. Histogram.

J-3. Box-and-Whiskers Plots.

J-3.1. *Introduction.* A box-and-whiskers plot (or a box plot) is a schematic diagram useful for visualizing important statistical quantities such as the center, spread, and distribution of a data set.

J-3.1.1. A box-and-whiskers plot (Figure J-5) is composed of a central box divided by a line and two lines extending out from the box called whiskers. The length of the central box—the interquartile range (IQR) or the distance from the 25th to the 75th percentile—indicates the spread of the bulk of the data (the central 50%) while the length of the whiskers shows the extent of the tails in the distribution. The length of each whisker is 1.5 IQR (roughly equal to two standard deviations for a normal data set). The width of the box has no particular meaning; the plot can be made quite narrow without affecting its visual impact. The sample median is displayed as a solid horizontal line through the box and the sample mean is displayed using a dotted horizontal line.

J-3.1.2. Box-and-whisker plots are useful for identifying possible outliers as they identify values that would be unusually large or small data if the data were assumed to be normally distributed. Any data points falling outside of the whiskers are displayed as "outliers" by an "o" or "x" on the plot. In particular, points falling 3.0 × IQR from the top or bottom of the box are "extreme outliers" displayed by an "x," while points falling 1.5 × IQR (but within 3.0 × IQR) from the top or bottom of the box are "mild outliers" displayed by an "o." For example, the box plot of the lognormal data in Figure J-5 contains three data values that are identified as unusual (two "mild outliers" and one "extreme outlier") if the data were assumed to be from a normal distribution. Each of the features described in this paragraph has been labeled in Figure J-5 to help you identify the most important features of box plots.

J-3.1.3. A box-and-whiskers plot can also be used to assess the symmetry of the data. If the distribution is symmetrical, then the box is divided in two equal halves by the median, the whiskers will be the same length and the number of extreme data points will be distributed equally on either end of the plot. For instance, the box plot of the normal data in Figure J-5 displays a highly symmetrical distribution of data. The mean and median are about the same, the 25th and 75th percentiles are about the same distance from the median, and the whiskers are roughly the same length. In contrast, the box plot of the lognormal data in Figure J-5 shows a noticeable positive skew. The mean is greater than the median, the upper whisker appears longer than the lower whisker, and several unusually large values are present on the upper end of the distribution. To see the variety in plots, the reader is urged to plot project-specific data.

J-3.1.4. Box-and-whiskers plots are extremely useful for visual comparisons of data from multiple sources when they are presented side-by-side. For example, separate box plots can be constructed for comparing background concentrations to site concentrations. This provides simultaneous comparison of the medians and IQRs of several sets of data. Another example where box plots can be useful is when trying to determine if an assumption of equal variances is valid, by qualitatively comparing the IQRs of two data sets (Appendix M). Directions for generating a box-and-whiskers plot are contained in Paragraph J-3.2 and an example follows in Paragraph J-3.3.

J-3.2. *Directions for Generating a Box-and-Whiskers Plot.*

J-3.2.1. Set the vertical scale of the plot based on the maximum and minimum values of the data set. Select a width for the box plot keeping in mind that the width is only a visualization tool. Label the width *W*; the horizontal scale then ranges from –½*W* to ½*W*.

J-3.2.2. Compute the upper quartile ($x_{0.75}$, the 75th percentile) and the lower quartile ($x_{0.25}$, the 25th percentile). Compute the sample mean and median. Compute the interquartile range (IQR). (Refer to Appendix D to do these computations, as necessary.)

J-3.2.3. Draw a box through points ($-\frac{1}{2}W$, $x_{0.75}$), ($-\frac{1}{2}W$, $x_{0.25}$), ($\frac{1}{2}W$, $x_{0.25}$), and ($\frac{1}{2}W$, $x_{0.75}$). Draw a line from ($\frac{1}{2}W$, $x_{0.50}$) to ($-\frac{1}{2}W$, $x_{0.50}$) and mark point (0, $\bar{x}$) with (+).

J-3.2.4. Compute the upper end of the top whisker by finding the largest data value x_L less than $x_{0.75} + 1.5 \times \text{IQR}$. Draw a line from (0, $x_{0.75}$) to (0, x_L). Compute the lower end of the bottom whisker by finding the smallest data value x_S greater than $x_{0.25} - 1.5 \times \text{IQR}$. Draw a line from (0, $x_{0.25}$) to (0, x_S).

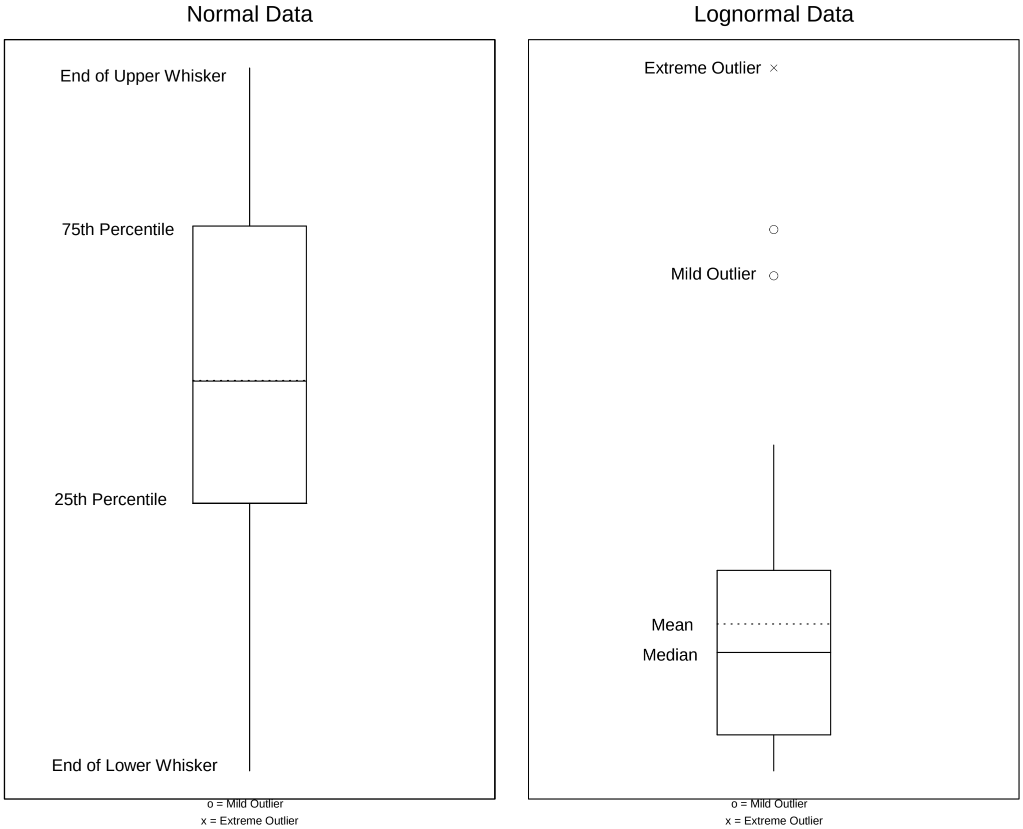


Figure J-5. Examples of box-and-whiskers plots.

J-3.2.5. For all points $x_L < x_* < x_{0.75} + 3.0 \times \text{IQR}$, place an “o” at the point $(0, x_*)$. These points are considered mild outliers. For all points $x_{**} > x_{0.75} + 3.0 \times \text{IQR}$, place an “x” at the point (0, x_{**}). These points are considered extreme outliers. Likewise, for all points $x_{0.25} - 3.0 \times \text{IQR} < x_{*} < x_{S}$, place an “o” at the point $(0, x_{*})$. Finally, for all points $x_{**} < x_{0.25} - 3.0 \times \text{IQR}$, place an “x” at the point $(0, x_{**})$.

J-3.3. *Example of a Box-and-Whiskers Plot*. Consider the following site data of chromium concentrations (mg/kg) in surface soil : 3.08, 3.35, 4.09, 4.13, 4.14, 4.36, 4.37, 4.42, 4.68, 4.76, 4.78, 4.82, 4.87, 4.89, 4.91, 4.94, 4.96, 4.96, 5.51, 6.4, 10.1, 10.3, 10.6 and 18.5

J-3.3.1. When generating the plot the width was set at a –0.25 to 0.25 horizontal range. Do not forget that the width is only a visualization tool and can be set to any value.

J-3.3.2. Compute the 75th percentile:

$$p = 0.75$$

$$np = 24 \times .75 = 18$$

$$np = j + g = 18 + 0$$

since $g = 0$

$$x_{0.75} = \frac{X_{(18)} + X_{(19)}}{2} = \frac{4.96 + 5.51}{2} = 5.235 \ .$$

J-3.3.3 Compute the 25th percentile:

$$p = 0.25$$

$$np = 24 \times .25 = 6$$

$$np = j + g = 6 + 0,$$

since $g = 0$

$$x_{0.25} = \frac{X_{(6)} + X_{(7)}}{2} = \frac{4.36 + 4.37}{2} = 4.365 \ .$$

Sample mean = 5.91, sample median = 4.845, interquartile range = $Q(.75) - Q(.25) = 0.87$.

J-3.3.4. Compute the upper end of the top whisker by finding the largest data value x_L less than

$$x_{0.75} + 1.5 \times IQR = 5.235 + 1.5(0.87) = 6.54\ .$$

So, x_L = 6.4. Draw a line from (0, 5.235) to (0, 6.4). Compute the lower end of the bottom whisker by finding the smallest data value x_S greater than

$$x_{0.25} - 1.5 \times IQR = 4.365 - 1.5(0.87) = 3.06\ .$$

So, x_S = 3.08. Draw a line from (0, 4.365) to (0, 3.08).

J-3.3.5. There are no points, x_*, greater than $x_L = 6.4$ but less than

$$x_{0.75} + 3.0 \times IQR = 5.235 + 3.0(0.87) = 7.845$$

so no points are considered mild outliers. For all points

$$x_{**} > x_{0.75} + 3.0 \times IQR = 7.845$$

place an "x" at the point (0, x_{**}). These points are considered extreme outliers. There are no points less than x_S = 3.08 so no points are drawn below the bottom whisker. Figure J-6 shows the box-and-whiskers plot.

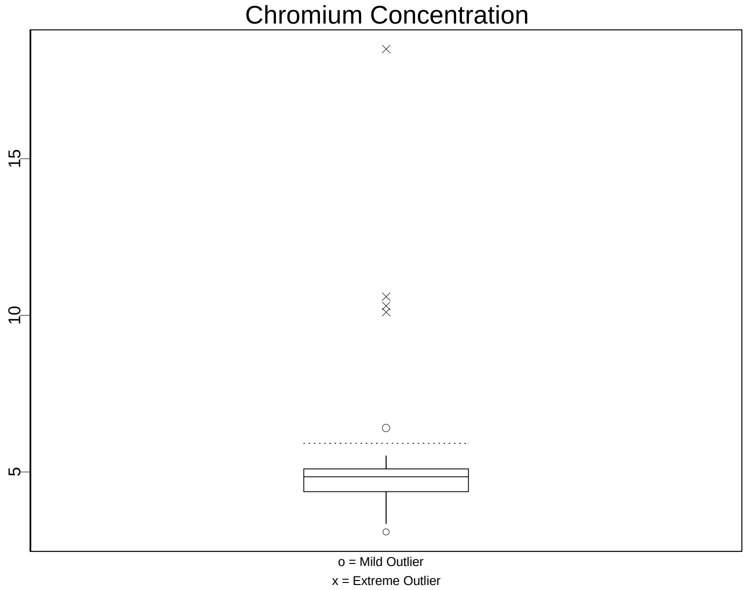


Figure J-6. Box-and-whiskers plot.

J-4. Quantile Plots.

J-4.1. *Introduction*. A quantile plot is a graph of the quantiles of data. It plots each point according to the fraction of the points it exceeds. It is a graphical representation of the data that is easy to construct, easy to interpret, and makes no assumptions about a model for the data.

J-4.1.1. A quantile plot displays every data point ranging from the lowest value to the highest value; it is a graphical representation of the data instead of a summary of the data. The advantage of using a quantile plot is that the analyst does not have to make any arbitrary choices regarding the data to construct a quantile plot (such as selecting the cell sizes for a making a histogram).

J-4.1.2. The vertical axis of the quantile plot is the measured concentration, and the horizontal axis of the quantile plot is the percentile of the data distribution. Directions for developing a quantile plot are given in Paragraph J-4.2 and an example follows in Paragraph J-4.3.

J-4.1.3. A quantile plot can be used to read quantile information (the median, quartiles, and the interquartile range) because each data value is plotted against the percentage of the data with that value or less. In addition, the plot can be used to determine the density of the data points: Are all the data values close to the center with relatively few values in the tails or are there a large number of values in one tail with the rest evenly distributed? The density of the data is displayed through the slope of the graph. A flat slope indicates a large number of data values; the graph rises slowly. A steep slope indicates a small number of data values; the graph rises quickly. A quantile plot can be used to determine if the data are skewed or if they are symmetrical. Figure J-7 shows examples of three quantile plots. If the data are symmetrical, then the top portion of the graph will stretch to the upper right corner in the same way the bottom portion of the graph stretches to the lower left, creating an s-shape similar to Figure J-7a. A quantile plot of data that are skewed to the right is steeper at the top right than the bottom left, as shown in Figure J-7b. A quantile plot of data that are skewed to the left increases sharply near the bottom left of the graph as shown in Figure J-7c.

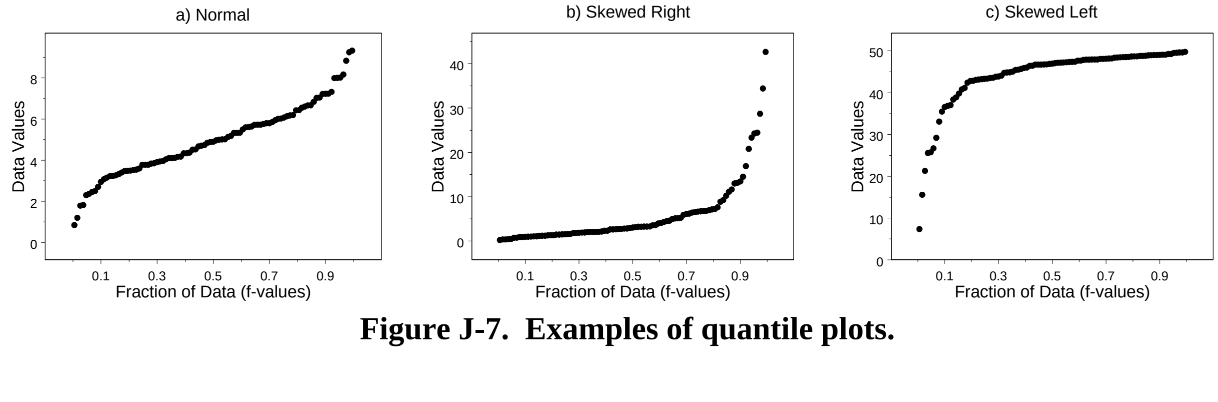


Figure J-7. Examples of quantile plots.

J-4.2. *Directions for Developing a Quantile Plot.* Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ represent the *n* data points ordered from least to greatest.

J-4.2.1. For each *i* from 1 to *n*, compute the fraction $f_i = (i - 0.5)/n$. The quantile plot is a plot of the pairs $(f_i, x_{(i)})$.

J-4.2.2. An example is given below in Paragraph J-4.3. (There are a number of ways to calculate the quantile f_i. Software that performs quantile plots may not necessarily use the same formula presented in Paragraph J-4.2 to calculate the quantile. For example, for the Weibull method $f_i = i/(n+1)$.)

J-4.3. *Generating a Quantile Plot.* Consider the following 108 data points for benzene groundwater results in µg/L: 0.0292, 0.0300, 0.0300, 0.0300, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0444, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0469, 0.0469, 0.0469, 0.0480, 0.0504, 0.0504, 0.0504, 0.0548, 0.0585, 0.0605, 0.0605, 0.0605, 0.0641, 0.0641, 0.0641, 0.0641, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0786, 0.0786, 0.0855, 0.0970, 0.0971, 0.1430, 0.2220, 0.2350, 0.3080, 0.4840, 0.6350, 0.7590, 0.8130, 1.1500, 1.7200, 1.7800, 1.8400, 1.8500, 1.9200, 2.0000, 2.0100, 2.1700, 2.1900, 2.3100, 2.4600, 2.6800, 2.7500, 2.9500, 3.4200, 3.4500, 3.7900, 4.3000, 5.4700, 5.7700, 5.8700, 6.1700, 6.9100, 7.2400, 7.5600, 8.3400, 8.6400, 9.3300, 11.0000, 11.1000, 12.2000, 14.1000, 17.0000, 20.2000, 21.8000, 29.1000, 36.7000 and 44.5000.

J-4.3.1. The data, ordered from smallest to largest, $x_{(i)}$, are shown in the first column of Table J-2 and the ordered number for each observation, *i*, is shown in the second column. The third column displays the values f_i for each *i* where $f_i = (\mathrm{i} - 0.5)/n$.

Table J-2.
Quantile Plot Data

$x_{(i)}$	i	f_i
0.0290	1	0.0046
0.0300	2	0.014
0.0300	3	0.023
.	.	.
.	.	.
.	.	.
29.100	106	0.9769
36.700	107	0.9861
44.500	108	0.9954

J-4.3.2. The pairs (f_i, x_i) are then plotted to yield the quantile plot in the Figure J-8.

J-5. Normal Probability Plots (Quantile-Quantile Plots). There are two types of quantile-quantile plots or *q-q* plots: an empirical quantile-quantile plot and a theoretical quantile-quantile plot. A normal probability plot is an extension of these *q-q* plots.

J-5.1. *Empirical Quantile-Quantile Plot*. A plot of the quantiles of two variables (e.g., the quantiles of *X* versus the quantiles of *Y*).

J-5.2. *Theoretical Quantile-Quantile Plot*. A plot of quantiles of a set of data against the quantiles of a specific theoretical probability distribution.

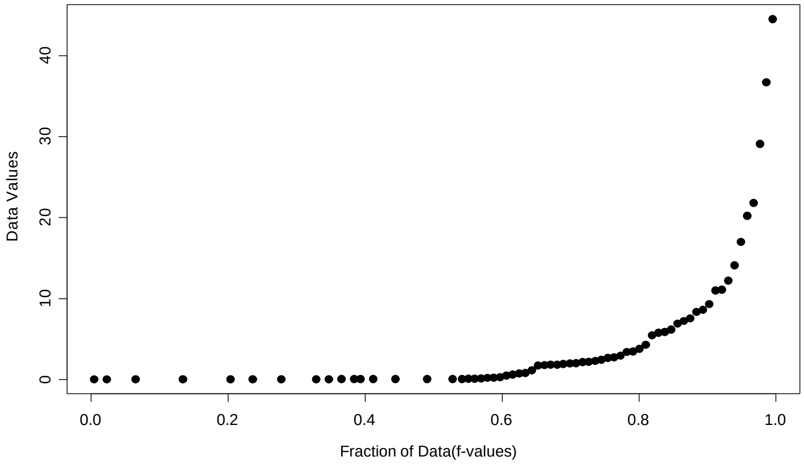


Figure J-8. Example of a quantile plot.

J-5.3. *Normal Probability Plot*. A theoretical quantile-quantile plot where the quantiles of a data set are plotted against the quantiles of the normal distribution.

J-5.4. *Introduction*. The following discussion will focus on the plot most commonly used for environmental data—the normal probability plot (the normal *q-q* plot); however, the discussion also holds for other *q-q* plots. The normal probability plot is used to roughly determine how well the data set is modeled by a normal distribution.

J-5.4.1. A normal probability plot, as defined above, is the graph of the quantiles of a data set against the quantiles of the normal distribution (see Figure J-9). If the graph is linear, the data may be normally distributed as shown in Figure J-9a. If the graph is not linear, the departures from linearity give important information about how the data distribution deviates from a normal distribution. Further, the graph may be used to determine the degree of symmetry (or asymmetry) displayed by the data. If the data in the upper tail fall above and the data in the lower tail fall below the quartile line, the data are too slender to be well modeled by a normal distribution (Figure J-9b); there are fewer values in the tails of the data set than what is expected from a normal distribution. If the data in the upper tail fall below and the data in the lower tail fall above the quartile line, then the tails of the data are too heavy to be well modeled using a normal distribution (Figure J-9c); there are more values in the tails of the data than what is expected from a normal distribution.

J-5.4.2. A normal probability plot can be used to identify potential outliers and extreme values. Data values much larger or much smaller than the rest will cause the other data values to be compressed into the middle of the graph, ruining the resolution. In addition, a normal probability plot is a useful technique for identifying irregularities in the data, especially in the tails, when compared to a certain distribution.

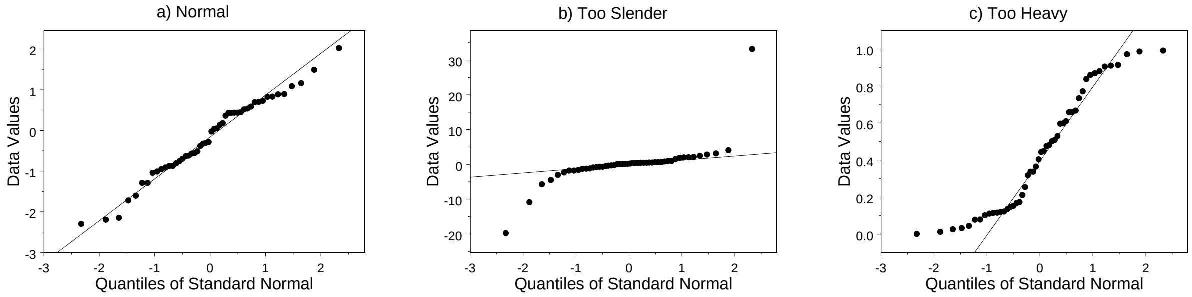


Figure J-9. Examples of normal probability plots.

J-5.4.3. Directions for constructing a normal probability plot are presented in Paragraph J-5.5, followed by an example in Paragraph J-5.6.

J-5.5. *Directions for Constructing a Normal Probability Plot*. Let $x_{(1)}$, $x_{(2)}$,..., $x_{(n)}$ represent the n data points ordered from least to greatest. For each i, compute the fraction $f_i = (i - 0.5)/n$ and find the corresponding quantile for the standard normal distribution, Z_p, in Table B-15 of Appendix B. The normal probability plot is a plot of the pairs (Z_p, $x_{(i)}$). If the data are normally distributed, the points will fall approximately on a straight line. The slope of the line is an estimate the population standard deviation and the y-intercept (at $Z = 0$) is an estimate of the population mean, because $X = \sigma Z + \mu$.

J-5.6. *Example for Constructing a Normal Probability Plot*. Again, consider the following results of benzene concentrations (in µg/L) in groundwater: 0.0292, 0.0300, 0.0300, 0.0300, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0444, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0469, 0.0469, 0.0469, 0.0480, 0.0504, 0.0504, 0.0504, 0.0548, 0.0585, 0.0605, 0.0605, 0.0605, 0.0641, 0.0641, 0.0641, 0.0641, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0786, 0.0786, 0.0855, 0.0970, 0.0971, 0.1430, 0.2220, 0.2350, 0.3080, 0.4840, 0.6350, 0.7590, 0.8130, 1.1500, 1.7200, 1.7800, 1.8400, 1.8500, 1.9200, 2.0000, 2.0100, 2.1700, 2.1900, 2.3100, 2.4600, 2.6800, 2.7500, 2.9500, 3.4200, 3.4500, 3.7900, 4.3000, 5.4700, 5.7700, 5.8700, 6.1700, 6.9100, 7.2400, 7.5600, 8.3400, 8.6400, 9.3300, 11.0000, 11.1000, 12.2000, 14.1000, 17.0000, 20.2000, 21.8000, 29.1000, 36.7000 and 44.5000.

J-5.6.1. The data, ordered from smallest to largest, are shown below in the first column of the table ($x_{(i)}$) and the ordered number for each observation (i) is shown in the second column. The third column displays the values f_i for each value of i, where $f_i = (i – 0.5)/n$. The fourth column displays the corresponding percentiles of the standard normal distribution, Z_p ($p = f_i$).

Table J-3.
Normal Probability Data

$x_{(i)}$	i	f_i	Z_p
0.0292	1	0.0046	–2.60
0.0300	2	0.014	–2.20
0.0300	3	0.023	–1.99
.	.	.	.
.	.	.	.
.	.	.	.
29.100	106	0.9769	1.99
36.700	107	0.9861	2.20
44.500	108	0.9954	2.61

J-5.6.2. The pairs (Z_p, $x_{(i)}$) are then plotted to yield the normal probability plot shown in Figure J-10. Because this plot is clearly nonlinear, these data are unlikely to be from a normal distribution.

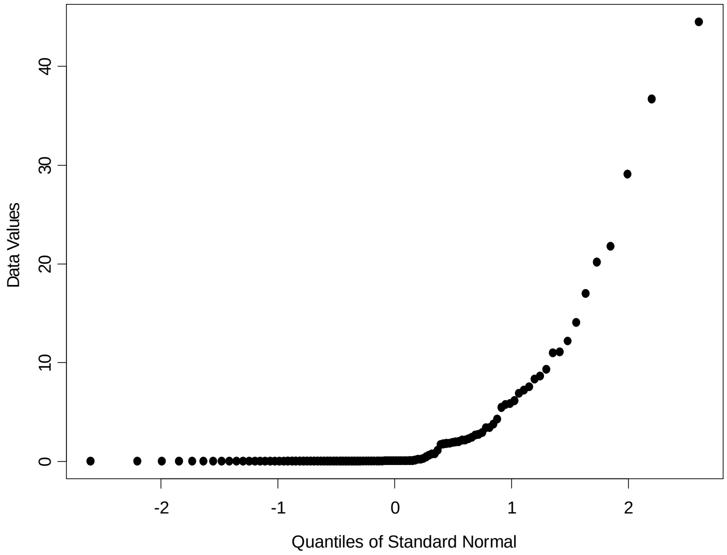


Figure J-10. Example of a normal probability plot.

J-6. Empirical Quantile-Quantile Plots. An empirical quantile-quantile (*q-q*) plot involves plotting the quantiles of two data variables against each other. This plot is used to compare distributions of two or more variables; for example, the analyst may wish to compare the distribution of lead and iron samples from a drinking water well. This plot is similar in concept to the theoretical quantile-quantile plot and yields similar information, plotting the distribution of two variables instead of the distribution of one variable in relation to a fixed distribution.

J-6.1. *Introduction.* If the distributions are roughly the same, the graph will be approximately linear; the slope will be nearly one and the intercept will be nearly zero. If the distributions are not the same, then the graph will not necessarily be linear. Even if the graph is not linear, the departures from linearity give important information about how the two data distributions differ. For example, a *q-q* plot can be used to compare the tails of the two data distributions in the same manner a normal probability plot is used to compare the tails of the data to the tails of a normal distribution. In addition, potential outliers (from the paired data) may be identified on this graph. Directions for constructing an empirical *q-q* plot are presented in Paragraph J-6.2 followed by an example in Paragraph J-6.3.

J-6.2. *Directions for Constructing an Empirical* q-q *Plot.* Let $x_1, x_2,..., x_n$ represent n data points of one variable and let $y_1, y_2,..., y_m$ represent a second variable of m data points.

J-6.2.1. Let $x_{(i)}$, for $i = 1$ to n, be the first sample listed in order from smallest to largest so that:

$x_{(1)}$ ($i = 1$) is the smallest

$x_{(2)}$ ($i = 2$) is the second smallest

…

$x_{(n)}$ ($i = n$) is the largest.

J-6.2.2 Let $y_{(i)}$, for $i = 1$ to m, be the second sample listed in order from smallest to largest so that:

$y_{(1)}$ ($i = 1$) is the smallest

$y_{(2)}$ ($i = 2$) is the second smallest

...

$y_{(m)}$ ($i = m$) is the largest.

J-6.2.3. If the two variables have the same number of observations, then an empirical *q-q* plot of the two variables is simply a plot of the ordered values of the variables. Because *n* = *m*, replace *m* by *n*. A plot of the following pairs is an empirical *q-q* plot:

$$(x_{(1)}, y_{(1)}), (x_{(2)}, y_{(2)}), ..., (x_{(n)}, y_{(n)}) .$$

J-6.2.4. If the two variables have a different number of observations ($n > m$), then the empirical *q-q* plot will consist of *m* (the smaller number) pairs. The empirical *q-q* plot will then be a plot of the ordered *y* values against interpolated *x* values. For *i* = 1, *i* = 2, ..., *i* = *m,*
let:

$$v = (n/m)(i - 0.5) + 0.5$$

and separate the result into the integer part and the fractional part, i.e., let:

$$v = j + g$$

where *j* is the integer part and *g* is the fraction part.

J-6.2.5. If *g* = 0, plot the pair ($y_{(i)}, x_{(i)}$). Otherwise, plot the pair $(y_i, (1-g)x_j + g\, x_{(j+1)})$. A plot of these pairs is an empirical *q-q* plot.

J-6.3. *Example for Constructing an Empirical* q-q *Plot.* Consider the following arsenic concentrations in subsurface soil samples (mg/kg): 2.15, 2.26, 2.37, 2.18, 1.93, 2.06, 2.00, 1.42, 1.31, 1.95, 2.88, 1.71, 1.92, 2.33, 1.55, 1.75, 2.09, 2.38, 2.11, 2.33, 1.98, 1.55, 1.76, 1.31, 2.34, 1.22, 1.81, 1.91, 2.31, 2.10, 1.89, 1.91, 1.49, 1.79, 2.71, 1.70, 1.93, 1.64, 1.94, 3.15, 2.32, 1.31, 1.97 and 1.48. And the following chromium concentrations in subsurface soil samples (mg/kg) are: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, 3.84, 2.95, 5.17, 4.80, 4.53, 4.01, 5.91, 3.96, 4.81, 5.27, 5.99, 4.60, 5.51, 4.72, 3.56, 4.22, 3.91, 5.81, 4.48, 5.10 ,4.94, 4.76, 4.62, 4.72, 4.73, 3.21, 4.14, 4.85, 4.25, 5.09, 3.68, 5.12, 6.60, 6.19, 3.15, 4.11 and 2.80.

J-6.3.1. An empirical *q-q* plot will be used to compare the distributions of these two analytes. As there are 44 observations of arsenic and 44 observations of chromium, the case for *m* = *n* will be used. Therefore, for *i* = 1, 2, ..., 44, compute:

$$(x_{(1)}, y_{(1)}), (x_{(2)}, y_{(2)}), ... (x_{(44)}, y_{(44)}) .$$

J-6.3.2. These pairs are plotted below, along with the best fitting regression line, as shown in Figure J-11.

J-7. Plots for Temporal Data.

J-7.1. *Introduction.* Data collected over specific time intervals (such as monthly, bi-weekly, or hourly) have a temporal component. For example, air monitoring measurements of a pollutant may be collected once a minute or once a day; water quality monitoring measurements of a contaminant level may be collected weekly or monthly. An analyst examining temporal data may be interested in the trends over time, correlation among time periods, or cyclical patterns, or all three. Some graphical representations specific to temporal data are the time plot, correlogram, and variogram.

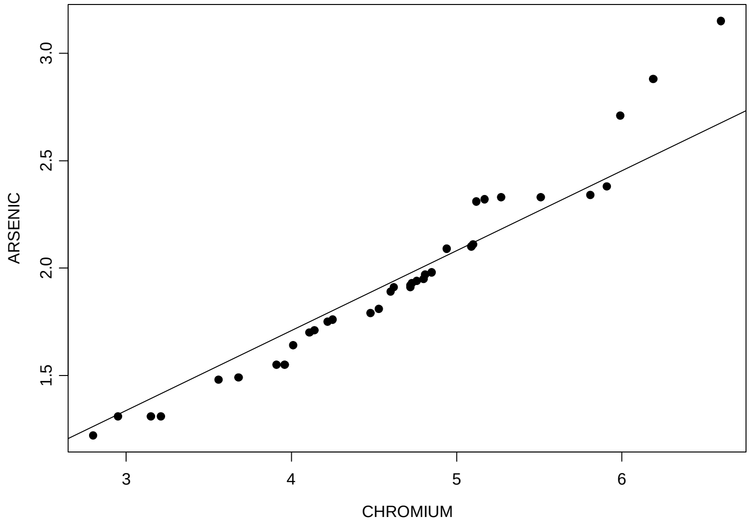


Figure J-11. Empirical *q-q* plot.

J-7.1.1. *Time Plot.* This is a plot of time versus some variable (e.g., concentration).

J-7.1.2. *Time Series Plot.* This is a time plot in which measurements of a variable are taken at regular, fixed intervals over time.

J-7.1.3. *Correlogram.* This is a plot that displays serial correlation when the data are collected at equally spaced time (or distance) intervals.

J-7.1.4. *Variogram.* This is a plot that displays the same information as a correlogram except that the data may be based on unequally spaced time (or distance) intervals. Further discussion of the variogram is contained in Appendix Q.

J-7.2. *Discussion.* Data collected at regular time intervals are called time series. The graphical representations presented in this Paragraph are recommended for all data that have a temporal component regardless of whether formal statistical time series analysis will be used to analyze the data. If the analyst uses a time series methodology or trend analyses such as those described in Appendix P, the graphical representations presented below will play an important role in this analysis. If the analyst decides not to use time series methodologies, these representations will help identify temporal patterns that need to be accounted for in the analysis of the data.

J-7.2.1. The analyst examining temporal environmental data may be interested in cyclic trends, directional trends, serial correlation, and stationarity.

J-7.2.1.1. *Cyclic Trend.* This is a pattern in the data (e.g., attributable to seasonal changes) that repeats over time.

J-7.2.1.2. *Directional Trend.* This is a downward or upward trend in the data.

J-7.2.1.3. *Serial Correlation.* This is a measure of the extent to which successive observations are related.

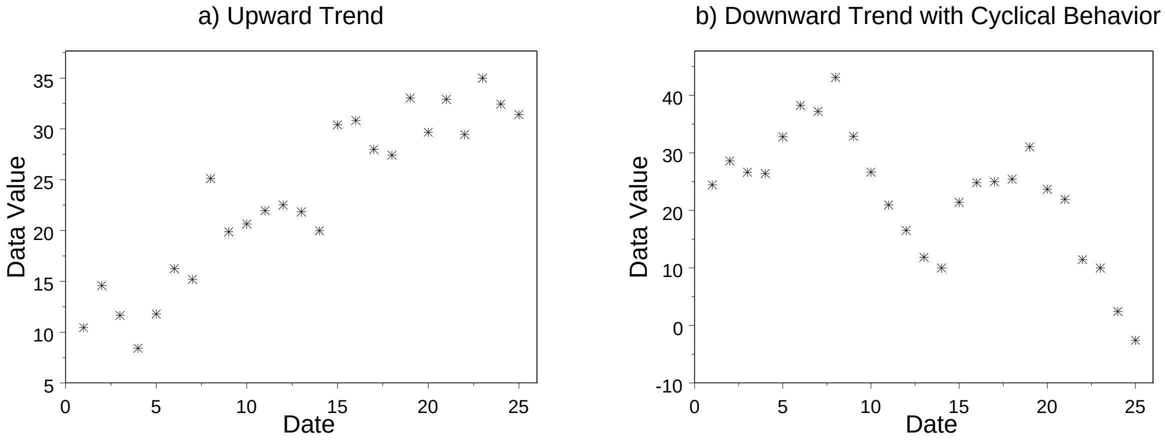
J-7.2.1.4. *Stationarity.* This describes the situation when the data looks the same over all time periods.

J-7.2.2. Cyclic trends repeat over time; the data rise and fall regularly over one or more time periods. These trends may be large scale, such as a yearly trend where the data show the same pattern of rising and falling over each year, or the trends may be small scale, such as a daily trend where the data show the same pattern for each day. Directional trends are downward or upward trends in the data, of importance to environmental applications where contaminant levels may be increasing or decreasing. Serial correlation is a measure of the extent to which successive observations are related. If they are related, statistical quantities calculated without accounting for it may be biased.

J-7.2.3. Another issue for temporal data is stationarity. Stationary data look the same over all time periods. Directional or cyclical trends and increasing (or decreasing) variability among the data imply that the data are not stationary. Temporal data are sometimes used in environmental projects along with a statistical hypothesis test to determine if contaminant levels have changed over time. If the hypothesis test does not account for temporal trends or seasonal variations, the data must achieve a steady state before the hypothesis may be tested. The data must be essentially the same for comparable periods of time both before and after the hypothesized time of change.

J-7.2.4. Sometimes multiple observations are taken in each time period. For example, the sampling design may specify selecting five samples every Monday for 3 months. If this is the case, the time plot may be used to display the data, display the mean weekly level, display a confidence interval for each mean, or display a confidence interval for each mean with the individual data values. A time plot of all the data can be used to determine if the variability for the different time periods changes. A time plot of the means can be used to determine if the means are changing between time periods. In addition, each time period may be treated as a distinct variable and the methods for plots for two or more variables may be applied.

J-7.3. *Time Plots*. One of the simplest plots to generate that provides a large amount of information is a time plot. This is a plot of the data that makes it easy to identify large- and small-scale trends over time. Small-scale trends show up on a time plot as fluctuations in smaller (or shorter) time periods. For example, ozone levels over the course of one day typically rise until the afternoon, then decrease, and this process is repeated every day. Larger scale trends, such as seasonal fluctuations, appear as regular rises and drops in the graph. For example, ozone levels tend to be higher in the summer than in the winter, so ozone data tend to show both a daily trend and a seasonal trend. A time plot can show directional trends and increased variability over time. Possible outliers may also be easily identified using a time plot. Figure J-12 displays two examples of time plots. Figure J-12a demonstrates an upward trend, while Figure J-12b shows a downward trend superimposed with cyclical behavior.


a) Upward Trend

b) Downward Trend with Cyclical Behavior

Figure J-12. Examples of time plots.

J-7.3.1. *Discussion*. A time plot is constructed by numbering the observations in order by time. The time ordering is plotted on the horizontal axis and the corresponding observation is plotted on the vertical axis. Although the points plotted on a time plot may be joined by lines, it is recommended that the plotted points *not* be connected to avoid creating a false sense of continuity. The scaling of the vertical axis of a time plot is of some importance. A wider scale tends to emphasize large-scale trends, whereas a smaller scale tends to emphasize small-scale trends. Using the ozone example above, a wide scale would emphasize the seasonal component of the data, whereas a smaller scale would tend to emphasize the daily fluctuations. Directions for constructing a time plot are contained in Paragraph J-7.3.2 along with an example.

J-7.3.2. *Directions for Generating a Time Plot.* Let x_1, x_2,..., x_n represent n data points listed in order by time, i.e., the subscript represents the ordered time interval. A plot of the pairs (i, x_i) is a time plot of this data.

J-7.3.2.1. Consider the following 15 benzene concentrations (µg/L) measured in groundwater (listed in order by day): 12.200, 3.790, 3.420, 5.470, 0.813, 1.840, 7.560, 4.300, 2.680, 6.170, 0.635, 2.190, 1.720, 1.150 and 0.484.

J-7.3.2.2. By labeling day 1 as 1, day 2 as 2, and so on, a time plot is constructed by plotting the pairs (i, x_i) where: i = the number of the day, and x_i = the concentration level.

J-7.3.2.3. A time plot of these data is shown in Figure J-13.

J-7.4. *Plot of the Autocorrelation Function (Correlogram).*

J-7.4.1. *Discussion.* Serial correlation is a measure of the extent to which successive observations are related. If successive observations are related, either the data must be transformed or this relationship must be accounted for in the analysis of the data. The correlogram is a plot that is used to display serial correlation when the data are collected at equally spaced time intervals. The autocorrelation function is a summary of the serial correlations of data. The first autocorrelation coefficient (r_1) is the correlation between all points that are one time unit (k_1) apart; the second autocorrelation coefficient (r_2) is the correlation between points that are two time units (k_2) apart; and so on. A correlogram (Figure J-14) is a plot of the sample autocorrelation coefficients in which the values of k versus the values of r_k are displayed.

J-7.4.1.1. The correlogram is used for modeling time series data and helps to determine if serial correlation is large enough to create problems in the analysis of temporal data using other methodologies. A quick method for determining if serial correlation is large is to place horizontal lines at ±2/n, where n is the number of samples on the correlogram (shown as horizontal lines on Figure J-14). Autocorrelation coefficients that exceed this value require further investigation.

J-7.4.1.2. In application, the correlogram is only useful for data at equally spaced intervals. To relax this restriction, a variogram may be used instead. The variogram displays the same information as a correlogram except that the data may be based on unequally spaced time (or distance) intervals. For more information on the construction and uses of the variogram, consult a statistician.

J-7.4.1.3. Directions for constructing a correlogram are contained in Paragraph J-7.4.2, followed by example calculations in Paragraph J-7.4.3.

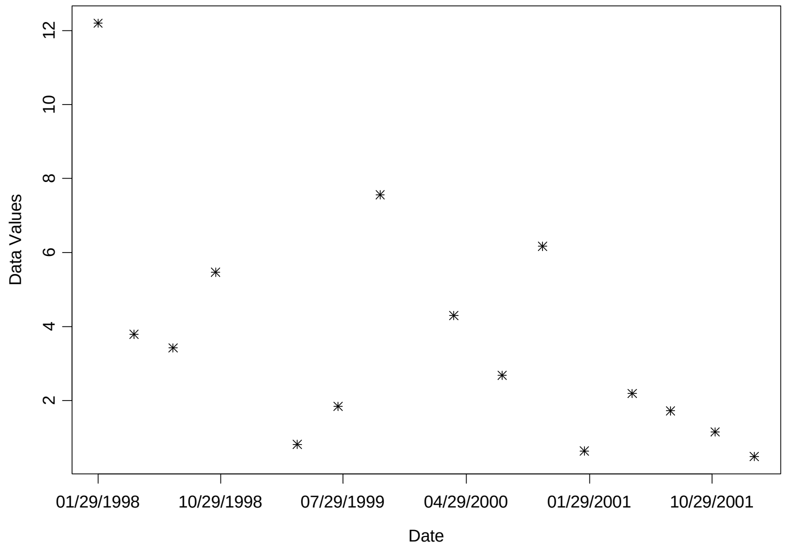


Figure J-13. Example time plot.

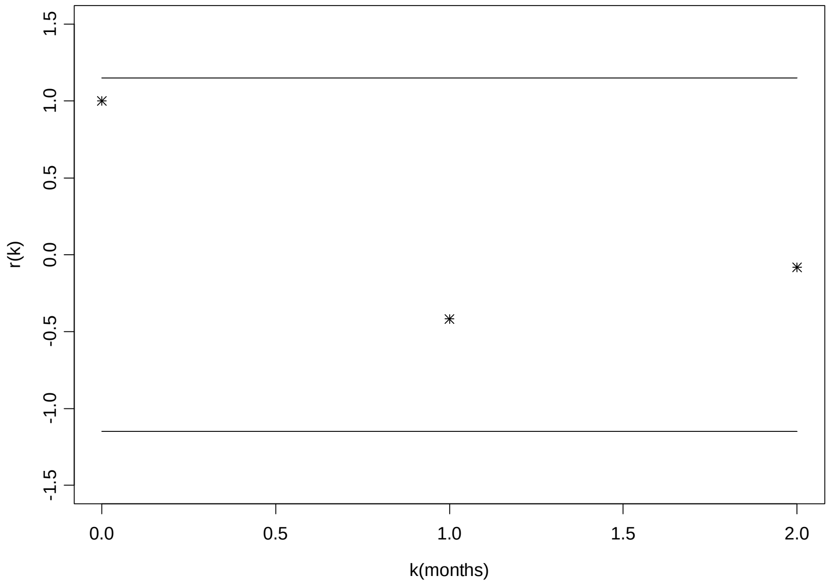


Figure J-14. Correlogram for data in Paragraph J-7.4.2.

J-7.4.2. *Directions for Constructing a Correlogram*. Let x_1, x_2,..., x_n represent the data points ordered by time for equally spaced time points, i.e., x_1 was collected at time 1, x_2 was collected at time 2, and so on.

J-7.4.2.1. To construct a correlogram, first compute the sample autocorrelation coefficients. So for $k = 0,1, \ldots..,$ compute r_k where

$$r_k = \frac{g_k}{g_o}$$

and

$$g_k = \frac{\sum_{t=k+1}^{n}(x_t - \bar{x})(x_{t-k} - \bar{x})}{n} \ .$$

J-7.4.2.2. Once r_k has been computed, a correlogram is the graph (r_k, k) for $k = 0, 1, \ldots$

J-7.4.2.3. Compute up to approximately $k = n/6$.

J-7.4.2.4. Also, note that $r_0 = 1$.

J-7.4.2.5. Finally, place horizontal lines at

$$\pm\frac{2}{\sqrt{n}} .$$

J-7.4.3. *Example for Constructing a Correlogram*. A correlogram will be constructed using the following three benzene concentrations in groundwater, collected monthly—month 1: 11.10 ppb, month 2: 2.46 ppb, month 3: 5.77 ppb. Although a correlogram would not typically be constructed when only three data points are available, only three data points are used here so that all computations may be shown. The rules that up to *n*/6 autocorrelation coefficients should be computed will be broken for illustrative purposes. The first step to constructing a correlogram is to compute the sample mean (Appendix D), which is 6.44 for the three points. Then,

$$g_0 = \sum_{t=1}^{3}(x_t - \bar{x})(x_{t-0} - \bar{x}) = \frac{\sum_{t=1}^{3}(x_t - \bar{x})^2}{3} = \frac{(11.10-6.44)^2 + (2.46-6.44)^2 + (5.77-6.44)^2}{3}$$

$$= \frac{21.72 + 15.84 + 0.45}{3} = 12.67$$

$$g_1 = \frac{\sum_{t=2}^{3}(x_t - 6.44)(x_{t-1} - 6.44)}{3} = \frac{(x_2 - 6.44)(x_1 - 6.44) + (x_3 - 6.44)(x_2 - 6.44)}{3}$$

$$= \frac{(2.46 - 6.44)(11.10 - 6.44) + (5.77 - 6.44)(2.46 - 6.44)}{3} = -5.29$$

$$g_2 = \frac{\sum_{t=3}^{3}(x_t - 6.44)(x_{t-2} - 6.44)}{3} = \frac{(x_3 - 6.44)(x_1 - 6.44)}{3} = \frac{(5.77 - 6.44)(11.10 - 6.44)}{3} = -1.05 \text{ .}$$

So,

$$r_1 = \frac{g_1}{g_0} = \frac{-5.29}{12.67} = -0.418$$

$$r_2 = \frac{g_2}{g_0} = \frac{-1.04}{12.67} = -0.082 \, .$$

Remember, r_0 = 1. Thus, the correlogram of these data is a plot of (0, 1) (1, –0.418) and (2, –0.082) with two horizontal lines at (±1.15). This graph is shown in Figure J-14. In this case, it appears that the observations are not serially correlated because all of the correlogram points are within the bounds of (±1.15).

J-7.4.4. *Multiple Observations Per Time Period.* In environmental data collection, multiple observations are sometimes taken for each time period. For example, the data collection design may specify collecting and analyzing five samples from a drinking well every Wednesday for 3 months. If this is the case, a time plot may be used to display the data, display the mean weekly level, display a confidence interval for each mean, or display a confidence interval for each mean with the individual data values. A time plot of all the data will allow the analyst to determine if the variability for the different collection periods changes. A time plot of the means will allow the analyst to determine if they are changing between the collection periods. In addition, each collection period may be treated as a distinct variable and the methods applied as described in the section on plots for two or more variables (Paragraph J-9).

J-8 Plots for Spatial Data.

J-8.1. *Introduction.* The graphical representations of the preceding Paragraphs may also be useful for exploring spatial data. An analyst examining spatial data may be interested in locating extreme values, overall spatial trends, and the degree of continuity among neighboring locations. Graphical representations for spatial data include postings, symbol plots, and correlograms (the correlograms would be generated by collecting samples at equally spaced sampling locations). The graphical representations presented below are recommended for all spatial data regardless of whether or not geostatistical methods will be used to analyze it. They will help identify spatial patterns that need to be accounted for in the analysis of the data. If geostatistical methods such as kriging are used to analyze the data, these methods will play an important role.

J-8.2. *Posting Plots*. A posting plot (Figure J-15) is a map of data locations along with corresponding data values. Data posting may reveal obvious errors in data location and identify data values that may be in error. The graph of the sampling locations gives the analyst an idea of how the data were collected (i.e., the sampling design), areas that may have been inaccessible, and areas of special interest to the decision-maker, which may have been heavily sampled. It is often useful to mark the highest and lowest values of the data to see if there are any obvious trends. If all of the highest concentrations fall in one region of the plot, the analyst may consider some method such as post-stratifying the data (stratification after the data are collected and analyzed) to account for this fact in the analysis. Directions for generating a posting of the data (a posting plot) are contained in Paragraph J-8.4.

J-8.3. *Symbol Plots*. For large amounts of data, a posting plot may not be feasible and a symbol plot (Figure J-16) may be used. A symbol plot is the same as a posting plot of the data, except that instead of posting individual data values, symbols are posted for ranges of the data values. For example, the symbol '0' could represent all concentration levels less than 100 ppm, the symbol '1' could represent all concentration levels between 100 ppm and 200 ppm, etc. Directions for generating a symbol plot are contained in Paragraph J-8.4.

J-8.4. *Directions for Generating a Posting Plot and Symbol Plot with an Example*.

J-8.4.1. *Directions*. On a map of the site, plot the location of each sample. At each location, either indicate the value of the data point (a posting plot) or indicate by an appropriate symbol (a symbol plot) the data range within which the value of the data point falls for that location, using one unique symbol per data range. The Posting plot and the Symbol plot are displayed as Figures J-15 and J-16.

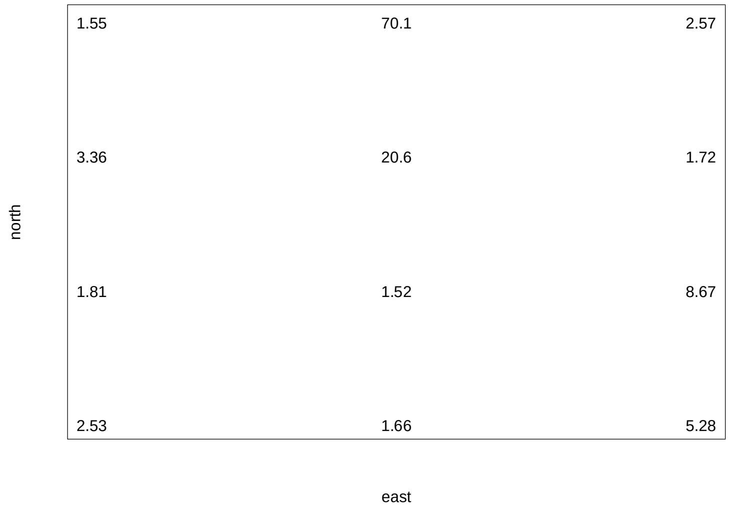


Figure J-15. Posting plot.

J-8.4.2. *Example*. The spatial data displayed in Table J-4 contains both a location (northing and easting) and a concentration level *C*. The data range from 4.0 to 35.5 so units of 5 were chosen to group the data.

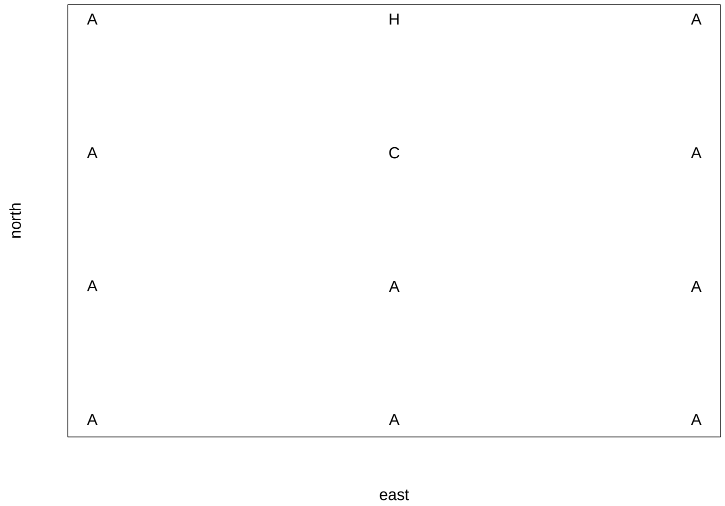


Figure J-16. Symbol plot.

Table J-4.
Spatial Data

Range	Symbol	Range	Symbol
0.0–9.9	A	40.0–49.9	E
9.9–19.9	B	50.0–59.9	F
20.0–29.9	C	60.0–69.9	G
30.0–39.9	D	70.0–79.9	H

Northing	Easting	*C*	Symbol	Northing	Easting	*C*	Symbol
25.0	0.0	2.53	A	15.0	15.0	2.57	A
25.0	5.0	1.81	A				
25.0	10.0	3.36	A				
25.0	15.0	1.55	A				
20.0	0.0	1.66	A				
20.0	5.0	1.52	A				
20.0	10.0	20.60	C				
20.0	15.0	70.10	H				
15.0	0.0	5.28	A				
15.0	5.0	8.67	A				
15.0	10.0	1.72	A				

J-8.5. *Other Spatial Graphical Representations.* The two plots discussed above, posting and symbol, provide information on the location of extreme values and spatial trends. The graphs below provide another item of interest to the data analyst, continuity of the spatial data. The graphical representations are not described in detail because they are mostly used for preliminary geostatistical analysis. These graphs can be difficult to develop and interpret. For more information on these, consult a statistician.

J-8.5.1. An *h* scatter plot is a plot of all possible pairs of data whose locations are separated by a fixed distance in a fixed direction (indexed by *h*). For example, an *h* scatter plot could be based on all the pairs whose locations are 1 meter apart in a southerly direction. An *h* scatter plot is similar in appearance to a scatter plot. The shape of the spread of the data in an *h* scatter plot indicates the degree of continuity among data values a certain distance apart in a particular direction. If all the plotted values fall close to a fixed line, then the data values at locations separated by a fixed distance in a fixed location are very similar. As data values become less and less similar, the spread of the data around the fixed line increases outward. The data analyst may construct several *h* scatter plots with different distances to evaluate the change in continuity in a fixed direction.

J-8.5.2. A correlogram is a plot of the correlations of the *h* scatter plots. Because the plot only displays the correlation between the pairs of data whose locations are separated by a fixed distance in a fixed direction, it is useful to have a graph of how these correlations change for different separation distances in a fixed direction. The correlogram is such a plot and allows the analyst to evaluate the change in correlation in a fixed direction as a function of the distance between two points. A spatial correlogram is similar in appearance to a temporal one. It spans opposite directions so that the correlogram with a fixed distance of due north is identical to the correlogram with a fixed distance of due south. Correlograms for spatial data are related to the semivariograms discussed in Appendix Q.

J-8.5.3. Contour plots are used to reveal overall spatial trends in the data by interpolating data values between sample locations. Most contour procedures depend on the density of the grid covering the sampling area (higher density grids usually provide more information than lower densities). A contour plot gives one of the best overall pictures of the important spatial features. However, contouring often requires that the actual fluctuations in the data values be smoothed, so that many spatial features of the data may not be visible. The contour map should be used with other graphical representations of the data and requires expert judgment to adequately interpret the findings.

J-9. Visualizing Higher Dimensional Data: Plots for Two or More Variables.

J-9.1. *Introduction.* To compare and contrast several variables, collections of the single variable displays described previously are useful. For example, the analyst may generate side-by-side box-and-whiskers plots or histograms for each variable using the same axis for all of the variables.

J-9.1.1. Figure J-17 illustrates side-by-side box-and-whiskers plots for naphthalene concentrations at various groundwater-monitoring wells at a given site.

J-9.1.2. In addition, the number of detected observations over the total number of observations has been placed towards the top of the graph. Separate plots for each variable may be overlaid on one graph, such as overlaying quantile plots for each variable on one graph. Another useful technique for comparing two variables is to place the histograms back to back. In addition, some special plots have been developed to display two or more variables; these allow comparison and contrast of individual data points of all the variables. These plots are described below.

J-9.2. *Plots for Individual Data Points.*

J-9.2.1. As it is difficult to visualize data in more than two or three dimensions, most of the plots developed to display multiple variables for individual data points involve representing each variable as a distinct piece of a two-dimensional figure. Such plots include Profiles, Glyphs, and Stars (Figure J-18). These graphical representations start with a specific symbol to represent each data point, then modify the various features of the symbol in proportion to the magnitude of each variable. The proportion of the magnitude is determined by letting the minimum value for each variable be of length zero, the maximum be of length one. The remaining values of each variable are then proportioned, based on the magnitude of each value in relation to the minimum and maximum.

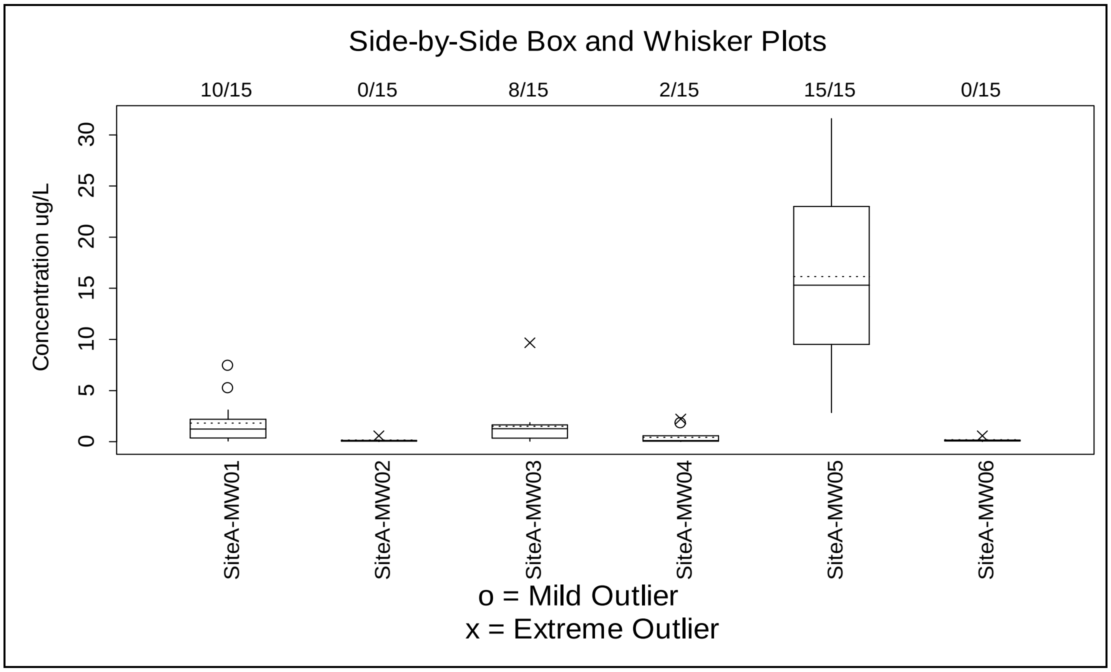


Figure J-17. Concentrations of naphthalene at site A wells.

J-9.2.2. A profile plot starts with a line segment of a fixed length. Then, lines spaced an equal distance apart and extended perpendicular to the line segment represent each variable. A glyph plot uses a circle of fixed radius. From the perimeter, parallel rays whose sizes are proportional to the magnitude of the variable extend from the top half of the circle. A star plot starts with a point where rays spaced evenly around the circle represent each variable and a polygon is then drawn around the outside edge of the rays.

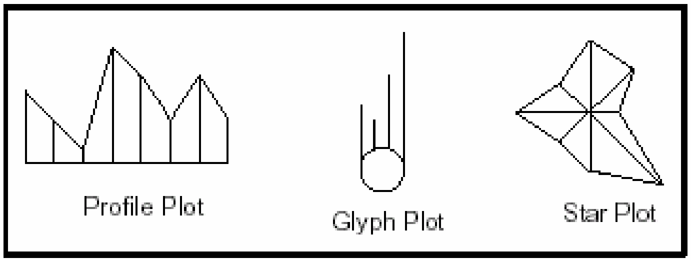


Figure J-18. Graphical representations of multiple variables.

J-9.3. *Scatter Plot.* For data sets consisting of paired observations where two or more continuous variables are measured for each sampling point, a scatter plot is one of the most powerful tools for analyzing the relationship between two or more variables. Scatter plots are easy to construct for two variables (Figure J-19) and many computer graphics packages can construct three-dimensional scatter plots. Directions for constructing a scatter plot for two variables are given in Paragraph J-9.4 along with an example in Paragraph J-9.5.

J-9.3.1. A scatter plot clearly shows the relationship between two variables. Both potential outliers from a single variable and potential outliers from the paired variables may be identified on this plot. A scatter plot also displays the correlation between the two variables. Scatter plots of highly linearly correlated variables cluster compactly around a straight line. In addition, nonlinear patterns may be obvious on a scatter plot. For example, consider two variables where one is approximately equal to the square of the other. A scatter plot of these data would display a U-shaped (parabolic) curve. Another important feature that can be detected using a scatter plot is any clustering effect among the data.

J-9.3.2. Additional information can be placed in a scatter plot. Labels can be placed on each value in the scatter plot to identify the sample location of a value. Different colors or symbols may be used to identify unique groupings of the data. For example, the scatter plot data may contain concentrations from multiple sampling events, with a unique symbol used to identify each event. This will show trends in concentrations as well as distinguishing sampling events.

J-9.4. *Directions for Generating a Scatter Plot.* Let x_1, x_2,..., x_n represent one variable of n data points and let y_1, y_2,..., y_n represent a second variable of the same n data points. The paired data can be written as (x_i, y_i) for i = 1,..., n. To construct a scatter plot, plot the first variable along the horizontal axis and the second variable along the vertical axis. It does not matter which variable is placed on which axis.

J-9.5. *Example of a Scatter Plot.* A scatter plot is prepared for arsenic and chromium concentrations in subsurface soil at Site A, using the data in Table J-5. Arsenic values are shown on the horizontal axis and chromium values are displayed on the vertical axis of Figure J-19.

J-9.6. *Extensions of the Scatter Plot.* It is easy to construct a two-dimensional scatter plot manually. Many software packages can construct useful two- and three-dimensional scatter plots. However, it is difficult to construct and interpret a scatter plot for more than three variables, so several graphical representations have been developed that extend the idea of a scatter plot to data consisting of two or more variables.

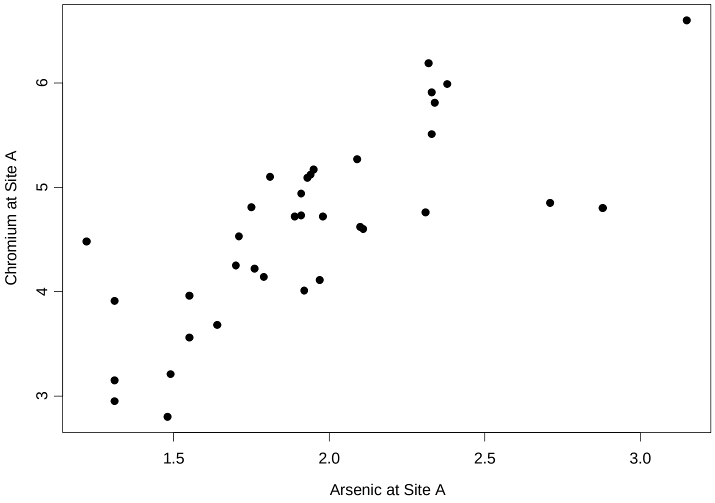


Figure J-19. Example of a scatter plot.

J-9.7. *Scatter Plot Matrix.* A scatter plot matrix is a useful method for extending scatter plots to higher dimensions. In this case, a scatter plot is developed for all possible pairs of the variables that are then displayed in a matrix format. This method is easy to use and is a concise method of displaying the individual scatter plots. However, this method does not contain information on three-way or higher interactions between variables. An example of a scatter plot matrix is contained in Figure J-20.

J-9.8. *Side-by-Side Scatter Plot.* A form of scatter plot, called a side-by-side scatter plot, is designed in a manner similar to the side-by-side box-and-whiskers plots presented earlier. Such scatter plots are developed using the horizontal axis as a label for each variable and using the vertical axis as the range of values for the variables. Figure J-21 illustrates a side by side scatter plots for the same data presented in Figure J-5. In Figure J-21, the *y*-axis is the range of concentrations for naphthalene and the *x*-axis represents the wells that were sampled during the site investigation. Because the wells were sampled over several years, different symbols are used to represent each year—triangles represent 1998, squares represent 1999, and circles represent 2000. In addition, because there are detected and non-detected results in the data, open symbols were used for non-detected values and closed symbols were used for detected values. At the top of the graph, a ratio is shown that states the number of detected observations over the total number of observations for each well sampled. A side-by-side scatter plot can be a useful tool in comparing and contrasting concentrations of a specific chemical at various data points (e.g., different wells at a particular site).

Table J-5.
Arsenic and Chromium Concentrations in Subsurface Soil at Site A

Sample ID	Arsenic (mg/kg)	Chromium (mg/kg)	Sample ID	Arsenic (mg/kg)	Chromium (mg/kg)
APA-EPC-SB01-030	1.31	2.95	APA-EPC-SB07-030	1.81	5.1
APA-EPC-SB01-040	1.95	5.17	APA-EPC-SB07-040	1.91	4.94
APA-EPC-SB01-050	2.88	4.8	APA-EPC-SB07-050	2.31	4.76
APA-EPC-SB02-030	1.71	4.53	APA-EPC-SB08-030	2.1	4.62
APA-EPC-SB02-040	1.92	4.01	APA-EPC-SB08-040	1.89	4.72
APA-EPC-SB02-050	2.33	5.91	APA-EPC-SB08-050	1.91	4.73
APA-EPC-SB03-030	1.55	3.96	APA-EPC-SB09-030	1.49	3.21
APA-EPC-SB03-040	1.75	4.81	APA-EPC-SB09-040	1.79	4.14
APA-EPC-SB03-050	2.09	5.27	APA-EPC-SB09-050	2.71	4.85
APA-EPC-SB04-030	2.38	5.99	APA-EPC-SB10-030	1.7	4.25
APA-EPC-SB04-040	2.11	4.6	APA-EPC-SB10-040	1.93	5.09
APA-EPC-SB04-050	2.33	5.51	APA-EPC-SB10-050	1.64	3.68
APA-EPC-SB05-030	1.98	4.72	APA-EPC-SB11-030	1.94	5.12
APA-EPC-SB05-040	1.55	3.56	APA-EPC-SB11-040	3.15	6.6
APA-EPC-SB05-050	1.76	4.22	APA-EPC-SB11-050	2.32	6.19
APA-EPC-SB06-030	1.31	3.91	APA-EPC-SB12-030	1.31	3.15
APA-EPC-SB06-040	2.34	5.81	APA-EPC-SB12-040	1.97	4.11
APA-EPC-SB06-050	1.22	4.48	APA-EPC-SB12-050	1.48	2.8

J-9.9. *Parallel Coordinate Plot.* A parallel coordinate plot also extends the idea of a scatter plot to higher dimensions. The parallel coordinates method employs a scheme where coordinate axes are drawn in parallel (instead of perpendicular). Consider a set of m-dimensional sample points $x_i = (x_{1i}, x_{2i}, x_{3i},..., x_{mi})$, where $i = 1, 2, 3…n$. For the i^{th} m-dimensional point, the variable $X_1 = x_{1i}$, $X_2 = x_{2i}$ and so forth. A parallel coordinate plot is constructed by first placing an axis (X_i) for each of the m variables parallel to each other. Each point x_{i} is graphically represented by plotting x_{1i} on the X_1 axis, x_{2i} on the X_2 axis and so forth, and then joining the set of m plotted values with a broken line. This method contains all of the information available on a scatter plot in addition to information on three-way and higher interactions (i.e., clustering among three variables). However, for m variables one must construct $m(m – 1)/2$ parallel coordinate plots in order to display all possible pairs of variables. For an example of a parallel coordinate plot see EPA QA/G-9 section 2.3.

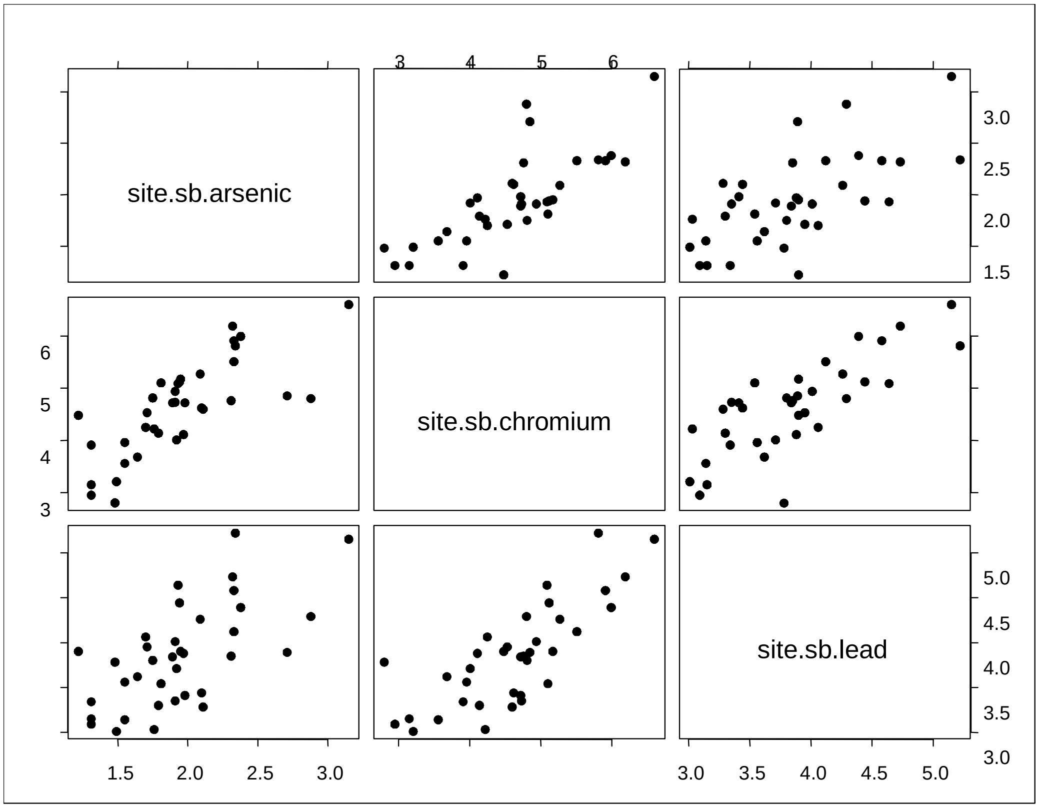


Figure J-20. Scatter plot matrix.

J-10. Contouring Data. Contouring site data helps in visualizing site conditions and presenting results. The results could be groundwater elevations and flow directions or locations and volumes of contamination. Contaminant concentrations are typically plotted by contouring the data over a site map. Contours or isopleths are lines of equal value (e.g., concentration). Lines or areas can be color coded or defined by a concentration range rather than a single value. Contour lines may not cross each other although they may form loops. The spacing of contour lines represents the gradient of the variable.

J-10.1. A topographic elevation map is a common contour map. Environmental data such as water table drawdown or chemical concentrations in water and air readily lend themselves to contouring. Contour maps are useful in data analysis because changes over distance, gradients, hot-spots, and the location of contaminants relative to site features, such as buildings and site boundaries, are apparent.

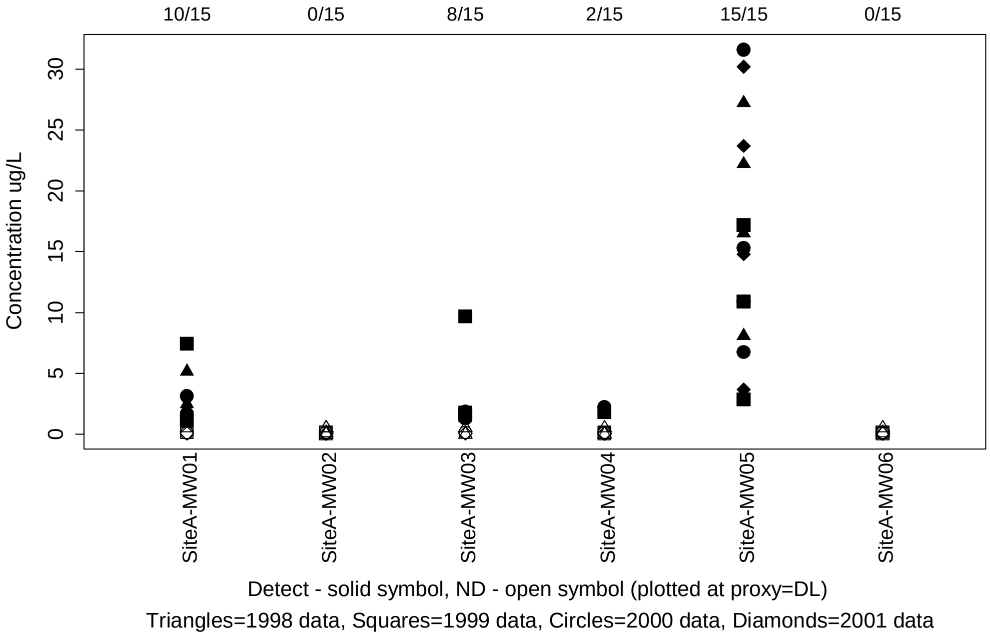


Figure J-21. Naphthalene concentrations at Site A wells: side-by-side scatter plots.

J-10.2. Contours must be interpolated because data coverage at a site is partial. For instance, water levels are measured only in monitoring wells even though the water table exists between wells. Interpolation estimates values within the existing data set while extrapolation estimates values outside the existing data set. Objects or values that are discontinuous spatially do not lend themselves to interpolation, for example, the presence of unexploded munitions at a test range.

J-10.3. There are numerous interpolation techniques available and selecting which to use depends on the media (soil, water, or air) and site-specific circumstances. Interpolation methods must be evaluated for their applicability, artifacts, and accuracy based on the analyst's site knowledge and technical expertise. Small data sets and software default values can result in contour maps that do not reflect actual site conditions. Software will often attempt to extrapolate beyond the data coverage unless a boundary is established or settings are carefully selected. It is good practice to contour data by hand and then compare results to computer-generated output. This allows the analyst to incorporate site-specific knowledge and intuition.

J-10.4. Currently available contouring software facilitates data interpretation and reinterpretation. Because data may be stored electronically, they may be readily revised and sorted. Numerous interpolation methods can be experimented with quickly. Pertinent information (such as sample depth, soil type, concentration) for a sampling point can be viewed by placing the cursor over it. High-end two-dimensional (2-D) and three-dimensional (3-D) color graphic images plus animation can be generated.

J-10.5. 3-D contouring, which is typically done with the aid of computers, is important to consider as an analysis tool. 3-D iso-surfaces are generated in lieu of 2-D contour lines. The failure to view contamination in natural systems in three dimensions, excluding the vertical or depth components, can adversely affect decision-making. It can give rise to misinterpretations of contamination sources (responsible parties) and transport, particularly when the geology is not laterally homogeneous or contaminants have densities different from the transport medium (solvents denser than water).

J-10.6. For field data to be adequately characterized, the manner in which the data will be analyzed should be considered when designing the sampling plan. The study area or area of concern should be well within the sample grid. This helps establish a boundary and ensure that measurements will be taken where they are most needed. The spacing of sampling locations affects the manner in which data will be analyzed and what can be learned from the data set. Poorly sized sampling grids can miss hot-spots or make the site appear more contaminated than it actually is. Poorly distributed data will lead to software drawing concentric contours around known values. It is often the case that vertical sample spacing is closer than the horizontal spacing. This situation can cause the vertical samples to unduly override the horizontal characteristics of the subsurface. Scaling features can be used to compensate for biased data sets.

J-10.7. No single interpolation method will be universally appropriate. In addition to trying more than one interpolation method, it is advisable to examine the computation used by the software. Some methods are better suited for certain data sets, such as those where values go from one extreme to another quickly or those where the changes are gradual and smooth. The mathematical function can also limit the interpolated value to a value not necessarily representative of site conditions. For example, simple inverse distance weighting (IDW) interpolates using the mean of two known values. The result is that the interpolated value lies between both known values and minimum and maximum values are not derived. It is also possible to interpolate negative values. Understanding the mathematical functions allows the input variables to be adjusted for individual circumstances. For instance, truncating a data set by setting a minimum and maximum concentration can alleviate some problems. The weight an interpolated point receives is directly related to its proximity to a known point. An interpolation's accuracy can be checked by randomly removing data points and then comparing the new interpolated value to the value that was removed.

J-10.8. Interpolation methods include the following.

J-10.8.1. *Linear Interpolation.* This is the mathematically simplest interpolation technique. This technique is referred to as manual or hand interpolation or contouring. A straight line drawn between two known values is subdivided into equal segments. The location of the estimated value is calculated using proportions.

J-10.8.2. *IDW Interpolation*. This gives more weight to an estimated point the closer it is to a known data point. IDW is often used for groundwater level data. A power value of two typically yields smooth contours.

J-10.8.3. *"Natural Neighbor" Interpolation*. This uses the same mathematical equation as IDW but the weighting technique is different. In addition, a polygon network is employed rather than a triangle network. The natural neighbor method may work well with clustered data.

J-10.8.4. *Triangular Irregular Network (TIN) Interpolation*. This connects the data points with a gridwork of triangles. TIN is used with linear interpolation to estimate values from the three vertices of each triangle.

J-10.8.5. *Spline Method*. This uses a polynomial function to fit a curve through the known points. It works well for data that change gradually. Splining is often applied to dense, regularly spaced data.

J-10.8.6. *Kriging*. This uses spatial variance to interpolate data. Kriging assumes, as IDW does, that distance and weight are related, but it also accounts for the spatial variance (spread) as a function of distance. Variograms, used in kriging, are graphs of a mathematical function that show spatial dependence in relation to distance and direction. Kriging has an intermediate step of matching the experimental variogram curve to a model variogram. Kriging handles steep gradients well, and is a good place to start for analyzing geological data because it was originally developed to predict ore locations for the mining industry. Variograms can provide insight into data sets even when kriging is not being performed.

J-10.9. Figure J-22 shows a groundwater elevation contour plot drawn by linear interpolation. Figure J-23 shows the same groundwater elevation data, factoring in the analyst's site knowledge. Figure J-24 shows the groundwater elevation data plot drawn by modeling software using IDW.

J-10.10. By using contouring with groundwater modeling software, otherwise static contour maps can be run forward or backward in time. This predictive modeling can be used to estimate the date at which some historical contaminant was released or plume migration at some future time. Groundwater modeling, GIS, statistics, and mapping software can perform various interpolation methods.

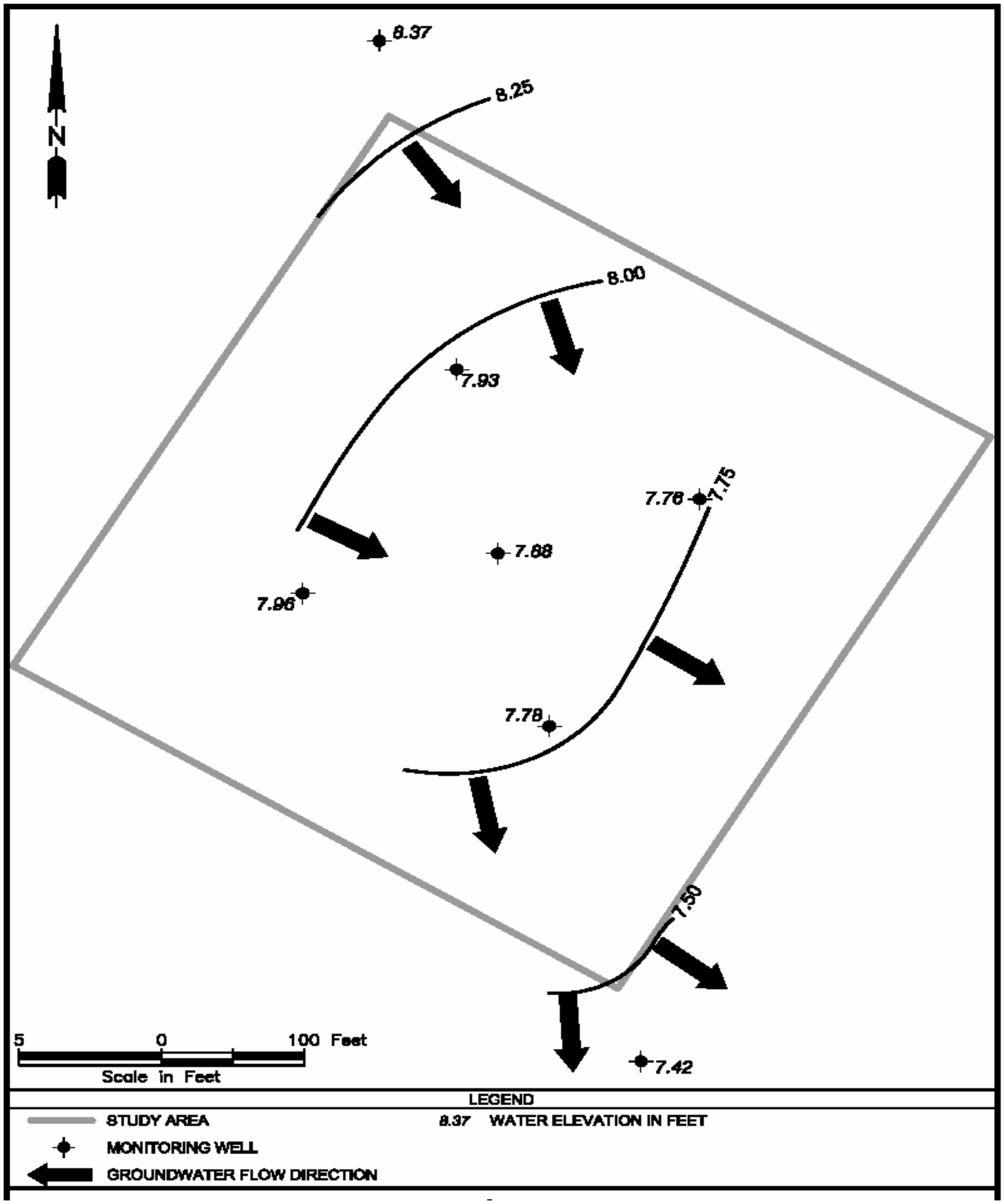


Figure J-22. Contour plot drawn by linear interpolation.

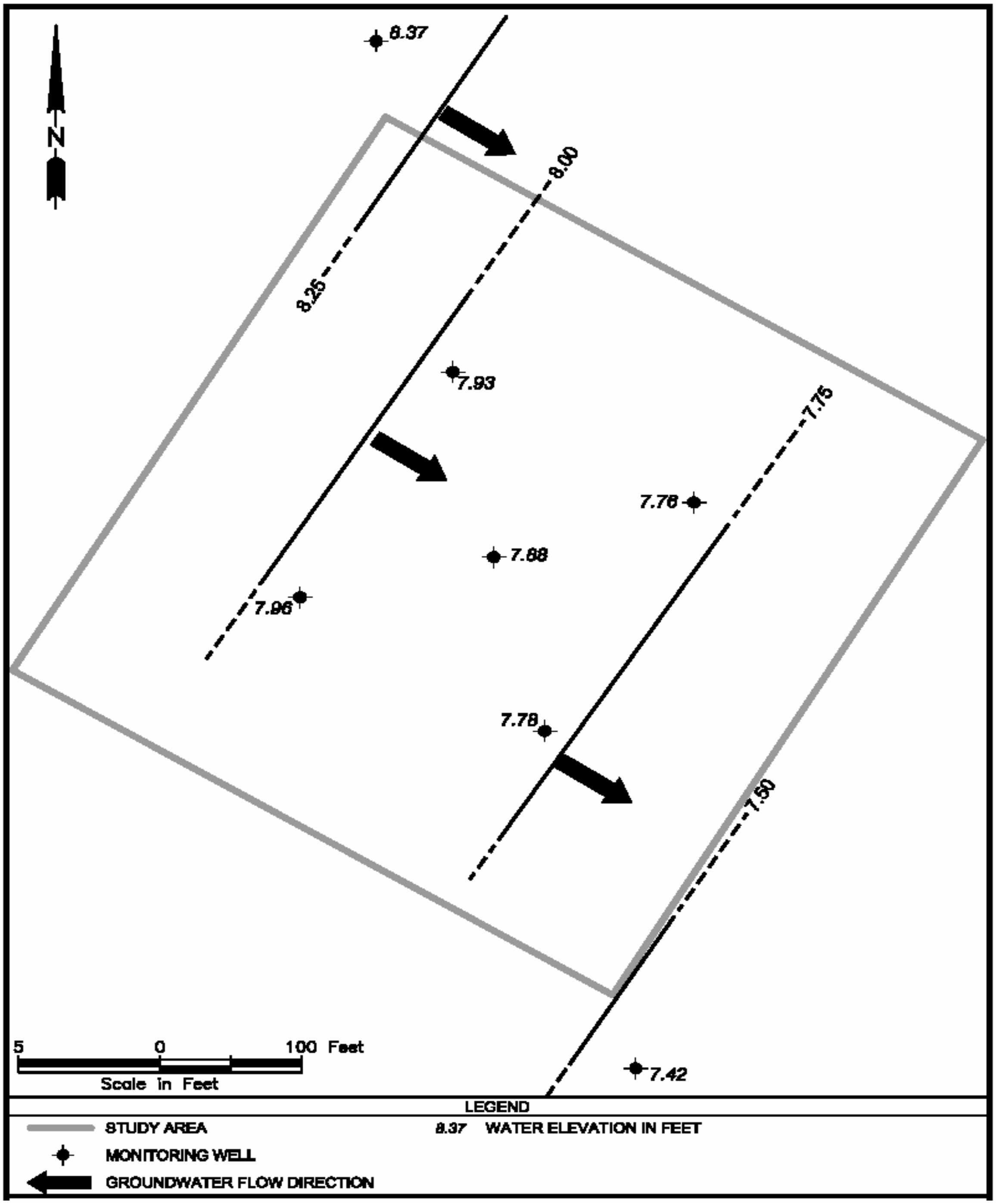


Figure J-23. Contour plot drawn by linear interpretation incorporating site knowledge.

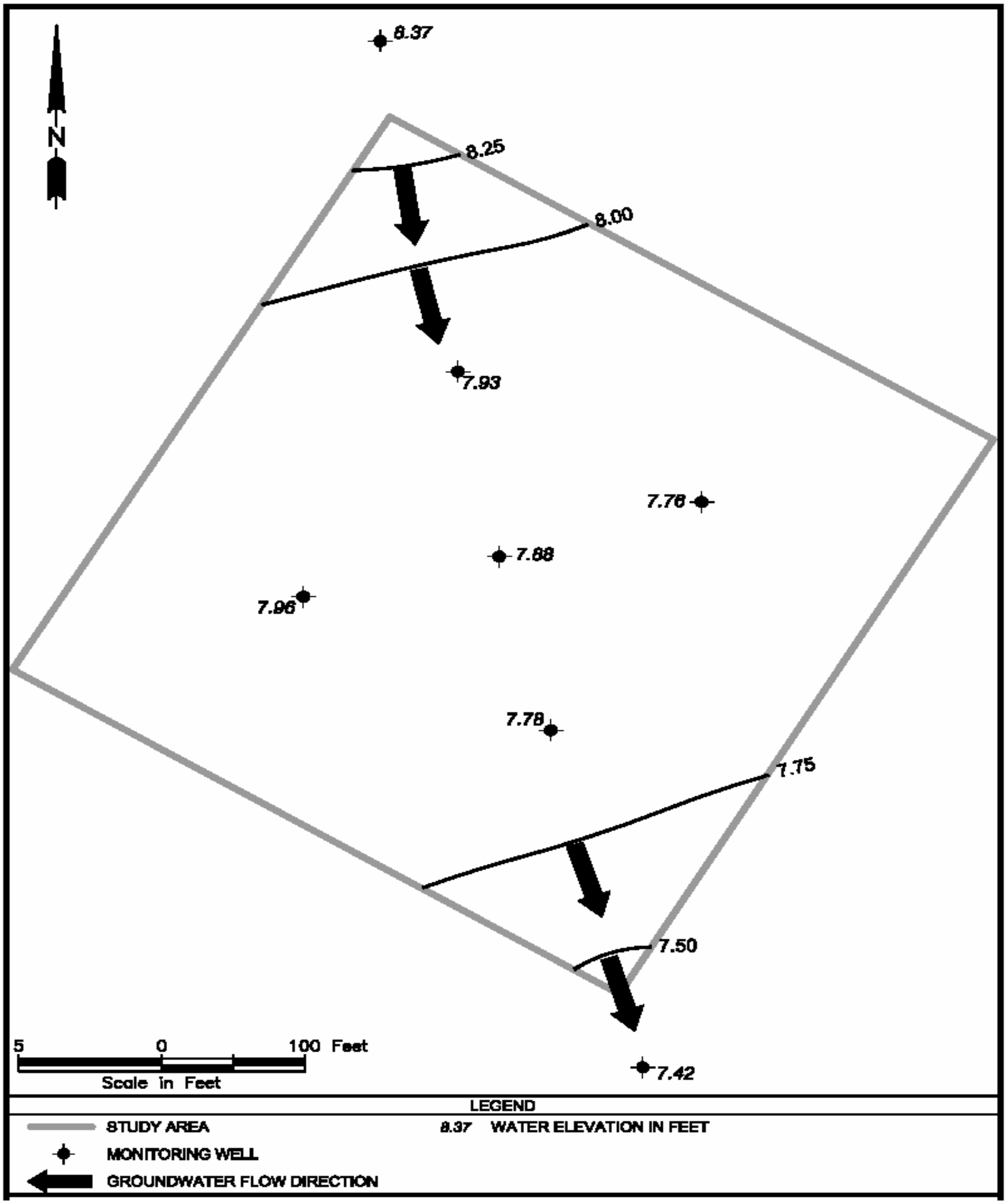


Figure J-24. Contour plot drawn by modeling software using IDW.