Walker Lake Data Set - An Exploratory Description

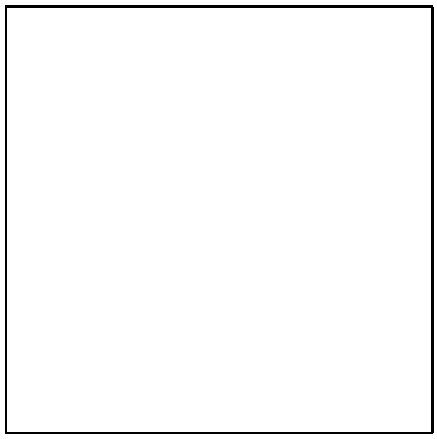
The text *An Introduction to Applied Geostatistics* uses a data set throughout to illustrate geostatistical concepts. This data set, known as the Walker Lake data set is summarized in what follows to illustrate some additional exploratory techniques. This data set is actually considered in 3 forms: an exhaustive data set, a smaller sample data set, and an even smaller illustrative data set. For each of these, two nameless variables, given only as U, V are used.

Exhaustive Data Set

- This is the full data set with 78000 sites on a 260x300 grid. This set will be viewed as the "population" of interest in the study.
- The variables U & V are *continuous* variables, both ranging from 0 to the thousands. The actual meaning of these variables is not given, but they are both viewed as concentrations (in ppm) throughout the text.
- This data set is primarily used as a means of testing whether or not analysis and subsequent inference on the smaller sample data sets were valid.

Sample Data Set

- This is a sample data set of the exhaustive set with 470 sites, with the same 2 variables.
- In addition to acting as a representative sample of the Walker Lake area, this sample data set illustrates some common problems encountered in data collection due to the sampling history. We will see later that initially 195 samples were collected on a regular grid where only U was measured. As the larger values of U were of greater interest, subsequent samples were taken in clusters near the sites with the largest U values. In this secondary sample, both U & V were measured. T was measured at all sites. Hence, there is missing V data at 195 of the 470 sites. This will make it very difficult to predict the value of V in areas where there was not heavy secondary sampling.



- This "double sampling" is a common sampling scheme, occurring when two variables U & V are related (as they are here), and U is cheaper to sample. The idea then is to first sample only U since it is inexpensive, find the areas of suspected large V values, and concentrate any sampling of V (which is expensive) in those areas.

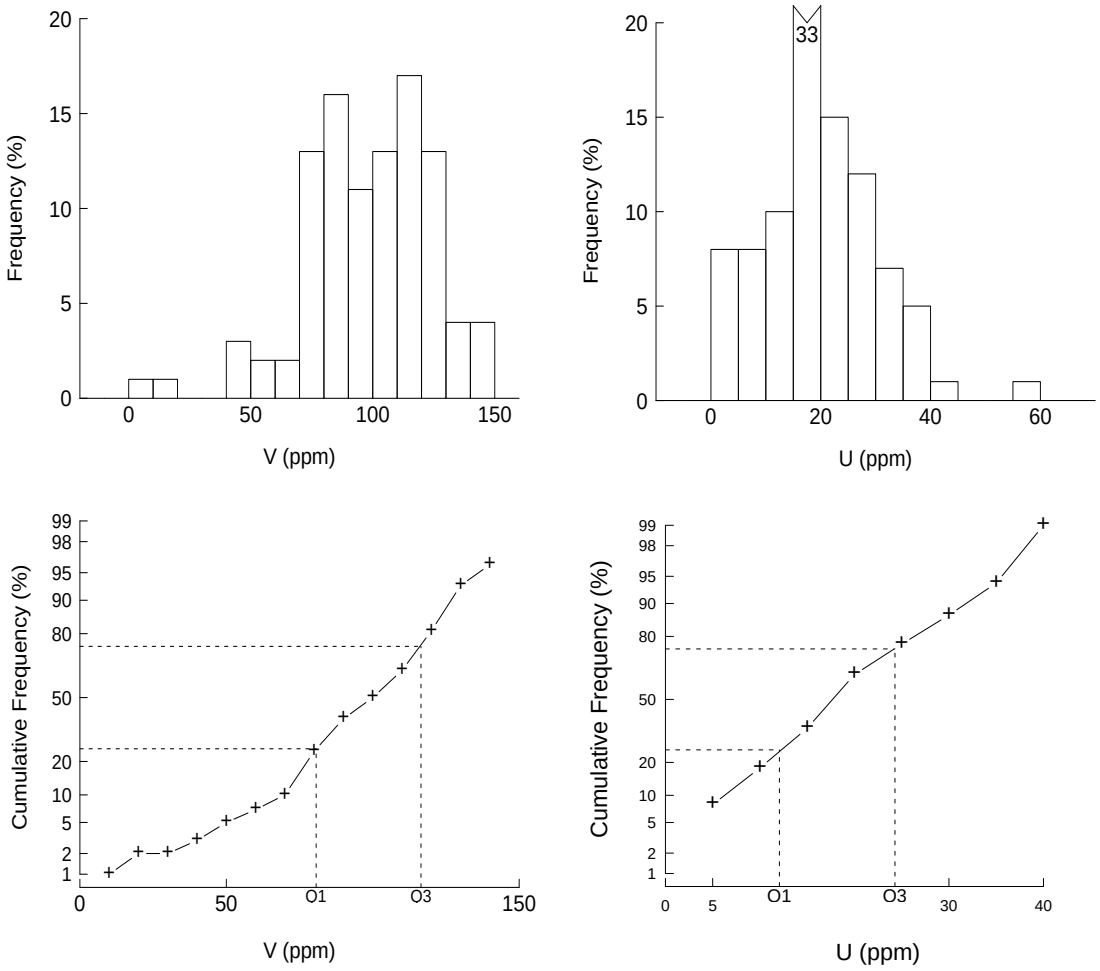
Illustrative Smaller Data Set

- This small data set is used to illustrate concepts, and consists of 100 sites on a 10x10 grid. The two variables U & V are measured on each site.

EDA Summary for the Illustrative (10x10) Data Set

Univariate Description

- The V data values look approximately normal (with 2 small values), as verified by both the histogram and normal probability plot shown below.
- The U data values look less normal with some right skewness, but probably do not deviate enough from normality to avoid using normal theory.



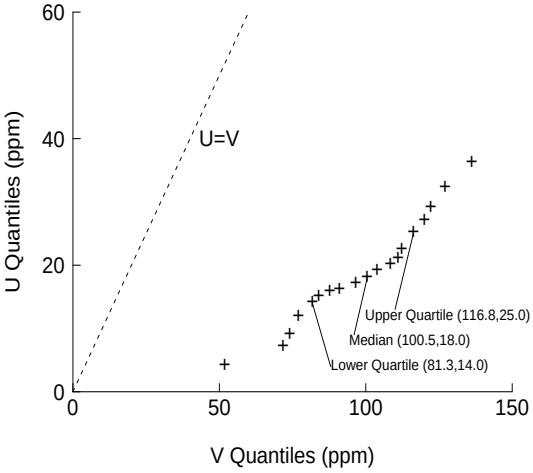
- One could also look at quantiles and other numerical summaries, such as the mean & median, standard deviation & IQR, and coefficients of skewness and variation.

Bivariate Description

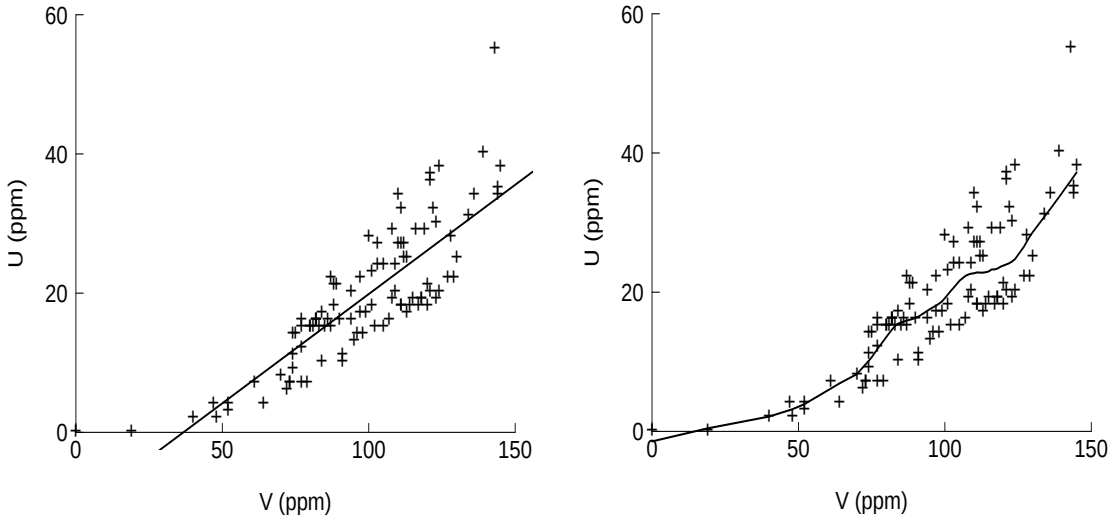
- A simple comparison of U & V values indicates that the V values are much larger and that both U and V exhibit some level of spatial correlation. The text plot, summary statistics, and a QQ-plot of the U quantiles versus the V quantiles show this very clearly, as given at the top of the next page.

81 + 15	77 + 12	103 + 24	112 + 27	123 + 30	19 + 0	40 + 2	111 + 18	114 + 18	120 + 18
82 + 16	61 + 7	110 + 34	121 + 36	119 + 29	77 + 7	52 + 4	111 + 18	117 + 18	124 + 20
82 + 16	74 + 9	97 + 22	105 + 24	112 + 25	91 + 10	73 + 7	115 + 19	118 + 19	129 + 22
88 + 21	70 + 8	103 + 27	111 + 27	122 + 32	64 + 4	84 + 10	105 + 15	113 + 17	123 + 19
89 + 21	88 + 18	94 + 20	110 + 27	116 + 29	108 + 19	73 + 7	107 + 16	118 + 19	127 + 22
77 + 15	82 + 16	86 + 16	101 + 23	109 + 24	113 + 25	79 + 7	102 + 15	120 + 21	121 + 20
74 + 14	80 + 15	85 + 15	90 + 16	97 + 17	101 + 18	96 + 14	72 + 6	128 + 28	130 + 25
75 + 14	80 + 15	83 + 15	87 + 15	94 + 16	99 + 17	95 + 13	48 + 2	139 + 40	145 + 38
77 + 16	84 + 17	74 + 11	108 + 29	121 + 37	143 + 55	91 + 11	52 + 3	136 + 34	144 + 35
87 + 22	100 + 28	47 + 4	111 + 32	124 + 38	109 + 20	0 + 0	98 + 14	134 + 31	144 + 34

	V	U
n	100	100
$\overline{x}$	97.6	19.1
s	26.2	9.81
CV	0.27	0.51
min	0.0	0.0
Q_1	81.3	14.0
M	100.5	18.0
Q_3	116.8	25.0
max	145.0	55.0



- U & V have a positive and fairly linear relationship. A least squares regression line fit through the data is shown below to the left, with a correlation coefficient of $r = 0.84$.



- If a simple linear regression model seemed inadequate for describing the relationship between U & V, there are nonparametric "smoothing" methods which can be used to estimate conditional expectation curves, using some form of local sliding neighborhoods. Two such popular methods are "loess smoothers" (locally weighted regression) and "smoothing splines." A loess fit is shown in the plot to the right above. One can adjust the degree of smoothness desired through the number of local points to use in the weighted regression.

- The **R** code and corresponding explanation for generating all of the plots in this handout can be found on the course webpage under the name **walker.r**. Additionally, the illustrative small Walker Lake data set can be found under the filename **walk100.txt** (text version). The data set consists of four columns with the x & y coordinates, v values and u values respectively.

Spatial Description: Here, some of the more common visual tools for studying spatial data are introduced, as well as means of identifying spatial trends and variance heterogeneity. Additionally, the notion of spatial continuity and the various types of stationarity assumptions made with spatial data are discussed.

Visual Tools

1. Data Posting: a map with the data values given in their spatial locations. Such plots are especially good for identifying outliers and large-scale trends. Two data postings of the V data values are given below, where the first highlights the lowest values and the second the highest values in the grid.

(a)

81	77	103	112	123	19	40	111	114	120
82	61	110	121	119	77	52	111	117	124
82	74	97	105	112	91	73	115	118	129
88	70	103	111	122	64	84	105	113	123
89	88	94	110	116	108	73	107	118	127
77	82	86	101	109	113	79	102	120	121
74	80	85	90	97	101	96	72	128	130
75	80	83	87	94	99	95	48	139	145
77	84	74	108	121	143	91	52	136	144
87	100	47	111	124	109	0	98	134	144

(b)

81	77	103	112	123	19	40	111	114	120
82	61	110	121	119	77	52	111	117	124
82	74	97	105	112	91	73	115	118	129
88	70	103	111	122	64	84	105	113	123
89	88	94	110	116	108	73	107	118	127
77	82	86	101	109	113	79	102	120	121
74	80	85	90	97	101	96	72	128	130
75	80	83	87	94	99	95	48	139	145
77	84	74	108	121	143	91	52	136	144
87	100	47	111	124	109	0	98	134	144

Figure 1: Location of the lowest V values in (a), and the highest values in (b).

2. Contour Maps: Contour maps are formed using interpolation algorithms and are good for spotting overall trends and the degree of change in the data. The contour map for the V data is shown to the right. Any observations?

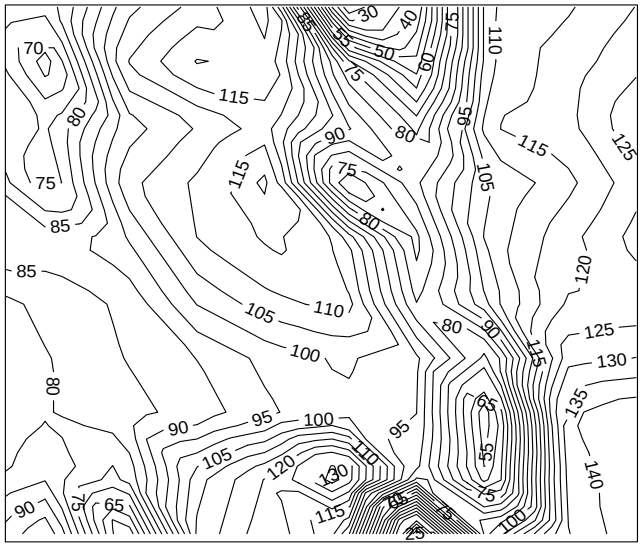


Figure 2: Computer generated contour map of 100 selected V data. The contour lines are at intervals of 10 ppm and range from 0 to 140 ppm.

3. Symbol Maps: the use of symbols to represent data, where each symbol represents a class of data values. The symbol plot for the V data is shown to the left below, where the numbers 0 through 9 are used to represent the range of V-values. It is generally easier to see trends or patterns in these types of plots than in data postings.

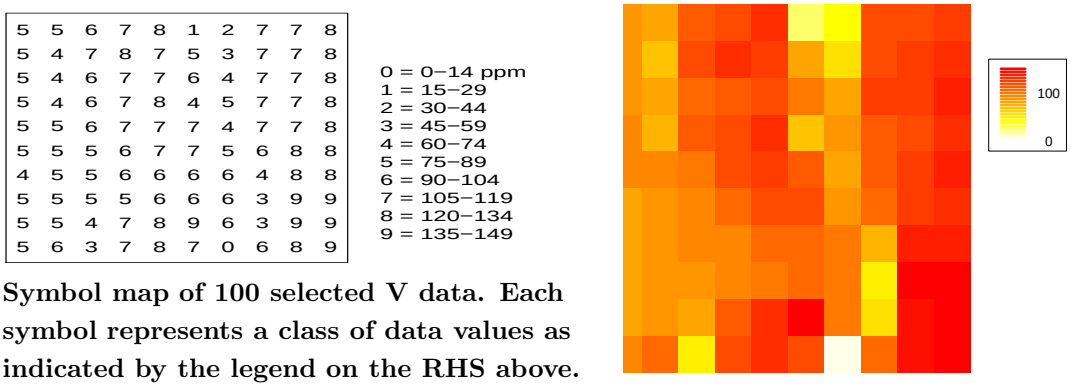


Figure 3: **Symbol map of 100 selected V data. Each symbol represents a class of data values as indicated by the legend on the RHS above.**

4. Grayscale Maps: maps shaded according to the size of the data values. A grayscale map for the V data is given to the right above.

5. Indicator Maps: symbol maps with 2 symbols (black, white). Such maps are useful for examining whether or not the data meet some threshold (i.e.: V $<$ 15 ppm vs. V $\geq$ 15 ppm). A series of indicator maps for the V data is given to the right. As the value of V increases, what do these plots indicate about patterns in the data?

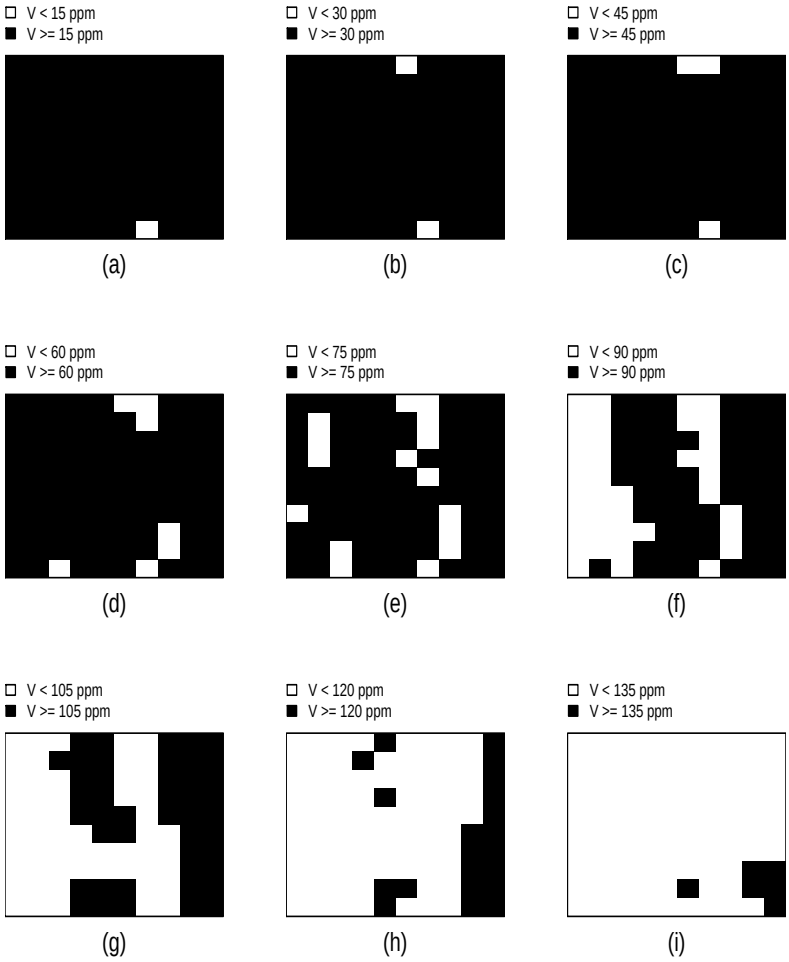


Figure 5: **Indicator maps of the 100 selected V data. Each figure is a map of indicators defined using the indicated cutoff values. The pattern of indicators provides a detailed spatial description of the data.**

A common question in geostatistical applications is whether or not the means and variances of the response variable V are the same across the study region D.

Unequal Means $\Leftrightarrow$

Unequal Variances $\Leftrightarrow$

- Typically in geostatistical data, there is no replication at the sites. So how do we estimate the mean and variance at different places in the region?
- Moving Window Statistics: Imagine placing a window (of any shape although typically a rectangle is used) in the region, where we compute the mean and variance using all data in the window, as it moves across the region. In this way, we get "local" estimates of the mean and variance for regions the size of the window, which we can then compare.

Issues:

1. Should the windows be overlapping or disjoint?

2. How many windows should we use? Is it better to have many small windows or fewer large ones?

Example: Consider the 10x10 grid of V-values from the illustrative data set in the text, shown in the two data postings below.

(a)

81	77	103	112	123	19	40	111	114	120
82	61	110	121	119	77	52	111	117	124
82	74	97	105	112	91	73	115	118	129
88	70	103	111	122	64	84	105	113	123
89	88	94	110	116	108	73	107	118	127
77	82	86	101	109	113	79	102	120	121
74	80	85	90	97	101	96	72	128	130
75	80	83	87	94	99	95	48	139	145
77	84	74	108	121	143	91	52	136	144
87	100	47	111	124	109	0	98	134	144

(b)

81	77	103	112	123	19	40	111	114	120
82	61	110	121	119	77	52	111	117	124
82	74	97	105	112	91	73	115	118	129
88	70	103	111	122	64	84	105	113	123
89	88	94	110	116	108	73	107	118	127
77	82	86	101	109	113	79	102	120	121
74	80	85	90	97	101	96	72	128	130
75	80	83	87	94	99	95	48	139	145
77	84	74	108	121	143	91	52	136	144
87	100	47	111	124	109	0	98	134	144

Figure 6: Example of overlapping moving windows for purposes of calculating moving average statistics.

- A 4x4 window was chosen to move throughout the grid (this gives 16 data values per window). The window was moved 2 meters at a time, giving 16 total windows.
- The mean and standard deviation was computed for the 16 values in each of the 16 windows, giving a 4x4 data posting of means and a 4x4 posting of variances, as shown in Figure 7 at the top of the next page.

- Observations?

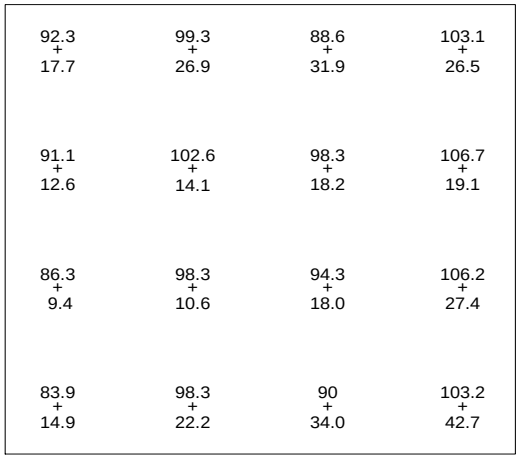
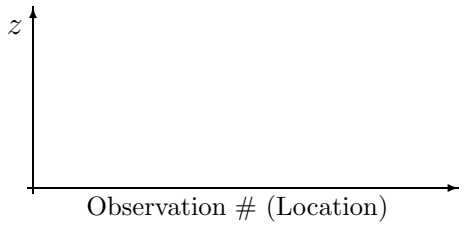


Figure 7: **Posting of statistics obtained from moving windows on the 100 V data. The mean of each moving window is plotted above the "+", and the standard deviation below.**

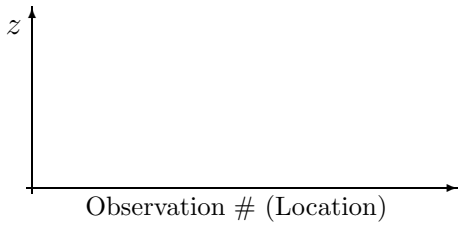
- In the presence of heteroskedasticity, what is typically done?

4 Common Relationships Between the Mean and Variance

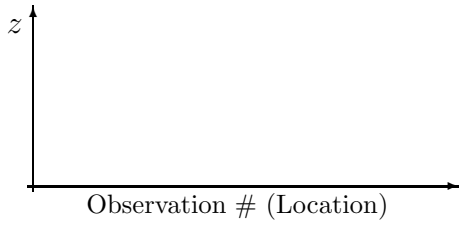
1. Mean & Variance are constant



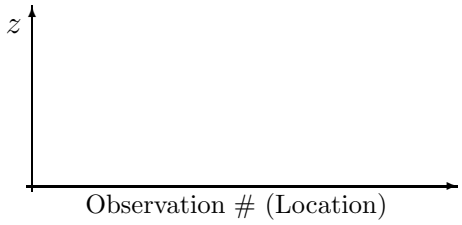
2.Mean varies, Variance is constant



3. Mean is constant, Variance varies



4. Mean & Variance both vary



To the right appears a plot (Figure 8) of the standard deviation s_V versus the mean m_V for the 16 pairs of means and standard deviations found using the 4x4 moving windows. Does there appear to be a relationship?

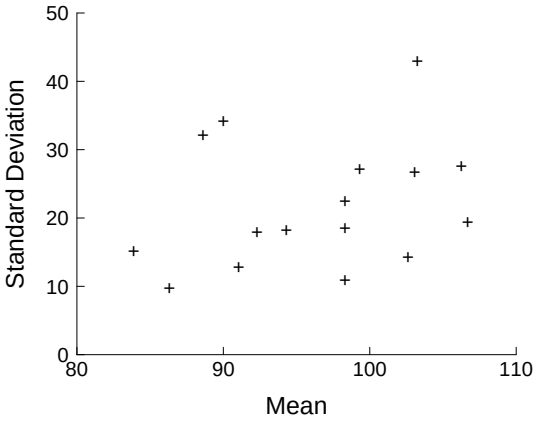


Figure 8: **Plot of the standard deviation vs. the mean obtained from moving windows on the 100 V data.**

- A relationship between the mean and standard deviation is called a proportional effect. Often, a transformation on the response variable Z can be performed to stabilize the variance:

  1. If $\sigma_i^2 \propto \mu_i$, try $Z' = \sqrt{Z}$ (common with Poisson count data).
  2. If $\sigma_i \propto \mu_i$, try $Z' = \log Z$ (common with concentration data).
  3. If $\sigma_i \propto \mu_i^2$, try $Z' = 1/Z$.
  4. If $z_{ij} = p_{ij}$ is a proportion, try $Z' = 2\text{arcsin}\sqrt{Z}$ (common with binomial count data).

Bottom Line: Many statistical methods either require variance homogeneity, or are greatly simplified when the variance is constant throughout the data. We will see that this is true of kriging methods in geostatistics as well.

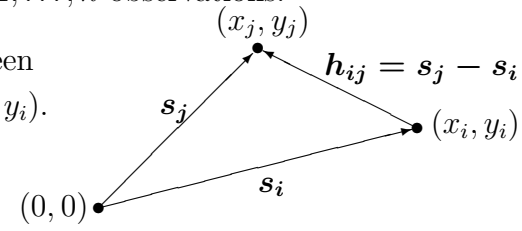
- Whether spatial trends exist or there is variance heterogeneity, most spatial data exhibit what is known as spatial continuity, the notion that sites close together are more likely to have similar data values than sites far apart. As it is on this premise that geostatistical methods are based, it is of interest to examine methods of assessing the degree of spatial continuity in a given set of data. How?

h-Scatterplots: An h-scatterplot is a visual tool used to examine the correlation between the response variable at a given site and the responses at neighboring locations.

Example: Suppose we are measuring soil moisture over a grid of points in a field where:

$$\begin{aligned} x &= \text{east-west coordinate, } y = \text{north-south coordinate,} \\ \boldsymbol{s_i} &= (x_i, y_i) = \text{the } i^{th} \text{ location, } i = 1, \ldots, n \text{ observations.} \end{aligned}$$

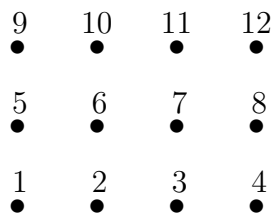
- Considering the two sites $\boldsymbol{s_i}, \boldsymbol{s_j}$, the vector between them is given by: $\boldsymbol{h_{ij}} = \boldsymbol{s_j} - \boldsymbol{s_i} = (x_j, y_j) - (x_i, y_i)$.



- Let Z_i be the response at location $\boldsymbol{s_i}$.
  Let Z_j be the response at location $\boldsymbol{s_j}$.

- The $\boldsymbol{h}$-Scatterplot is simply a scatterplot of z_j versus z_i for all pairs of locations separated in distance and direction by $\boldsymbol{h}$. Note that there is no ij subscript on $\boldsymbol{h}$ here because $\boldsymbol{h}$ is the same for any two locations separated by the same distance and same orientation.

Example: Suppose we have the 12 locations shown to the right. Note that the numbering of these sites is completely arbitrary.



- If we want the correlation in responses at neighboring sites in an E-W direction, we use an $\boldsymbol{h}$=(1,0)-scatterplot.

  Resulting Pairs:

- If we want the correlation in responses at neighboring sites in an E-W direction separated by 2 units, we use an $\boldsymbol{h}$= (2,0)-scatterplot. Resulting Pairs?

- Which plot do you expect to have a stronger relationship? Why?

- In general, h-scatterplots can be done in any direction and at any distance. This is clear for data on a regularly-spaced grid of points, but what if the data are not in the form of a grid?

Example: Consider again the V-concentration values on a 10x10m grid.

For an $\boldsymbol{h}$=(0,1)-scatterplot, how many pairs are there?

For an $\boldsymbol{h}$=(0,2)-scatterplot, how many pairs are there?

- In viewing the series of h-scatterplots shown, it should be noted that as the distance of separation increases, the relationship between $V(\boldsymbol{t})$ & $V(\boldsymbol{t}+\boldsymbol{h})$ becomes weaker.

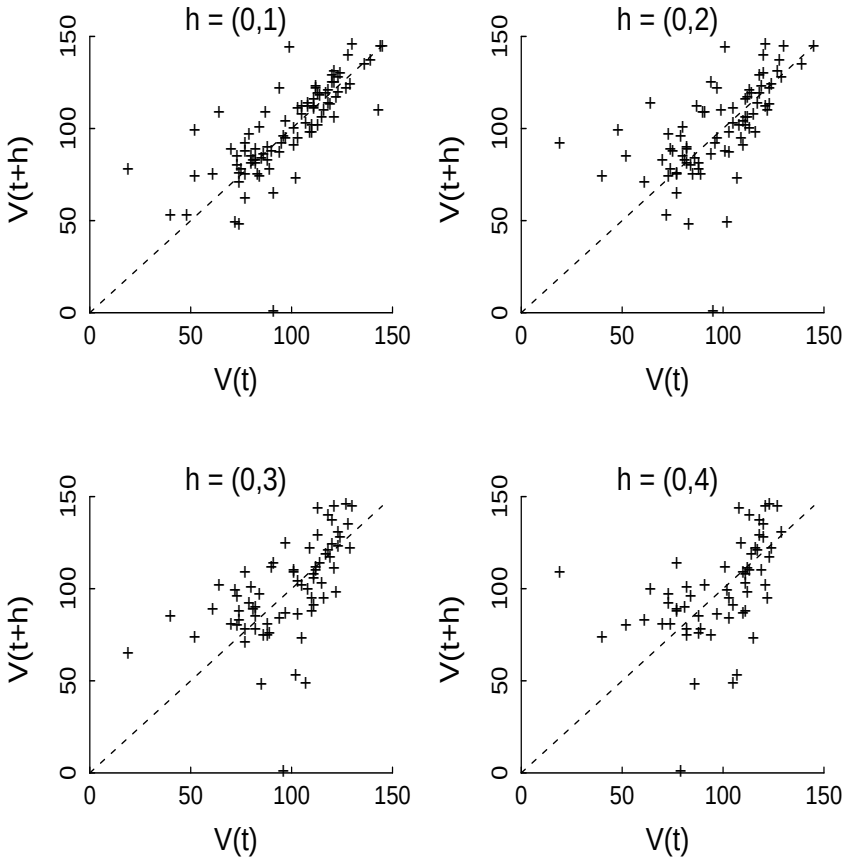


Figure 9: **h-Scatterplots for four separation distances in a northerly direction between pairs of the 100 V values. As the separation distance increases, this similarity between pairs of values decreases and the points on the h-scatterplot spread out further from the diagonal line.**

- How do we describe this spatial dependence numerically?

These functions summarize the strength of association between "neighboring" responses. Before defining these terms formally, we need to consider some basic assumptions commonly made on the mean and covariance structure of a spatial process.

Covariance Function, Correlogram, Variogram

In the previous handout, a visual tool known as the h-scatterplot was introduced to examine the strength of association between observations of the response variable as a function of distance and direction. The covariance function, correlogram, and variogram (or semivariogram) are all functions that numerically characterize the strength of such associations. With most spatial data, there are two common assumptions that are made.

1. Spatial Continuity: The spatial autocorrelation between the responses at 2 sites only depends on the distance and perhaps direction of orientation, not on where these sites are located in the region of interest.

2. Stationarity: Additionally, it is often assumed that the mean and variance are constant across the region of interest (see readings: pp. 55-60 (I&S), pp. 32-35 & p. 162 (B&G)).

In both time series analysis and spatial analysis, it is necessary to make some kind of assumption such as these in order to estimate the correlation pattern and the variance because realizations of the data cannot be assumed independent. Both of these two assumptions essentially allow for global homogeneity, so that different parts of the region can be treated as if they were replicates. This enables computation of a *common* covariance function for all parts of the region of interest.

There are two basic types of stationarity assumptions, outlined below. The first is known as covariance, second-order, or weak stationarity, and the second as intrinsic stationarity.

1. **Covariance** or **Second-Order** or **Weak Stationarity**: This is assumed for the covariance function (and corresponding correlation function or correlogram).
   - $E(Y_i) = E(Y_j)$ for all sites $i, j \in \mathcal{R}$. The mean is constant over the region $\mathcal{R}$.
   - $\text{Cov}(Y_i, Y_j) = C(\boldsymbol{h})$, $\boldsymbol{h} = (s_{1i}, s_{2i}) - (s_{1j}, s_{2j})$. This implies that the variance is the same everywhere and the covariance between two response variables depends only on the distance and direction between the two sites, not the location.

2. **Intrinsic Stationarity**: This is assumed for the variogram (or semivariogram).
   - $E(Y_i) = E(Y_j)$ for all sites $i, j \in \mathcal{R}$. The mean is constant over the region $\mathcal{R}$.
   - $\frac{1}{2}\text{Var}(Y_i - Y_j) = \gamma(\boldsymbol{h})$, $\boldsymbol{h} = (s_{1i}, s_{2i}) - (s_{1j}, s_{2j})$. This implies that the variance of the *difference* is the same everywhere. The variance of the Y_i's may not be the same everywhere.

Which of these two assumptions is weaker?

These stationarity assumptions are not absolutely necessary to conduct spatial analyses, but are required for many of the types of analyses we will consider. An active area of research concerns spatial analysis with nonstationary covariance structures.

Having examined the assumptions made for the various types of numerical functions which characterize spatial autocorrelation, these functions are now introduced.

Covariance Function (Covariogram)

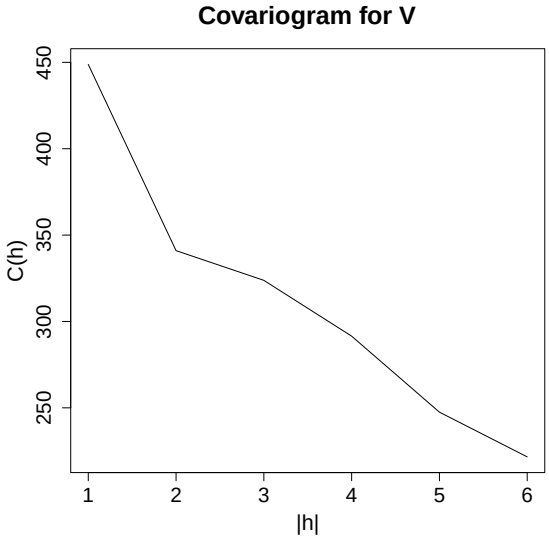
The covariogram summarizes the information given in h-scatterplots, and is a function of both the distance and direction between two locations in the region.

Notation: We use the notation $C(\cdot)$ and $\gamma(\cdot)$ to represent the *population* covariogram and variogram, respectively. These functions will be estimated using a sample covariogram and sample variogram, represented by $\widehat{C}(\cdot)$ and $\widehat{\gamma}(\cdot)$, respectively.

Example: Consider the V concentration data from the Walker Lake data set on a 10x10 grid. If we want to characterize the spatial autocorrelation in a north-south direction in this region, we could compute the covariance between all sites one unit apart in the N-S-direction, $\widehat{C}(0,1)$, two units apart in the N-S-direction, $\widehat{C}(0,2)$, ..., k units apart in the N-S-direction, $\widehat{C}(0,k)$. This set of covariances can then be plotted against the lag distance to examine the decrease in covariance with increased separation. A separate graph can be created for each direction desired.

A covariogram for the Walker Lake data in the N-S direction is given below along with a table of the corresponding covariances. These were found using the **hscatter** function in R.

$\boldsymbol{h}$	Covariance (ppm^2)
(0,1)	448.8
(0,2)	341.0
(0,3)	323.8
(0,4)	291.5
(0,5)	247.5
(0,6)	221.6



- Typically, we expect the covariance to be high for observations taken close together. At a distance of $h = 0$, we have:

  $C(h) =$

- Usually, we expect the covariance to decrease as units get further apart, so that eventually the units are uncorrelated and $\text{Cov}(Y_i, Y_j) = 0$ if the distance between $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$ is sufficiently large.

- The correlation function is called the correlogram. It is a standardized version of the covariance function which ranges from -1 to 1. We expect the correlation to be high for units close together (correlation = 1 at distance zero) and tend to zero as the separation distance between units increases.

- The theoretical covariance and correlation function for two random variables W and Z are given by:

  $\text{Cov}(W, Z) =$

  $\text{Corr}(W, Z) =$

- If we observe the bivariate sample $(W_1, Z_1), \ldots, (W_n, Z_n)$, the covariance above is estimated by the following sample covariance:

  $\widehat{\text{Cov}}(W, Z) = \widehat{\text{C}}(W, Z) =$

Sample Covariance Function (Covariogram)

Let: $\boldsymbol{s_i} = (s_{1i}, s_{2i}) = i^{th}$ location and

$\boldsymbol{h_{ij}} = \boldsymbol{s_j} - \boldsymbol{s_i}$, the vector connecting points $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$.

The sample covariogram is defined as: $\boxed{\widehat{C}(\boldsymbol{h}) = \frac{1}{n(\boldsymbol{h})} \sum_{(i,j)|\boldsymbol{h_{ij}} = \boldsymbol{h}} y_i y_j - m_{-\boldsymbol{h}} m_{+\boldsymbol{h}}}$, where:

$n(\boldsymbol{h})$ = the number of pairs $\boldsymbol{h}$ units apart

$m_{+\boldsymbol{h}}$ = the mean of all y_j values in the sum

$m_{-\boldsymbol{h}}$ = the mean of all y_i values in the sum

- The "y"s are made small in the above expression to indicate that they are realizations (or values) of the random variable Y.

- This is the sample covariance between Y at one location and Y at another location at a direction and distance away defined by $\boldsymbol{h}$.

- We are implicitly making the assumption that the covariance depends only on distance (& direction) and not on the particular location.

- It should be noted that $m_{-\boldsymbol{h}}$ and $m_{+\boldsymbol{h}}$ are in general not equal.

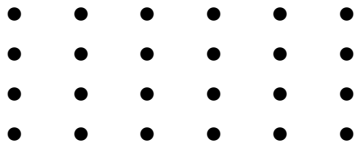
- Bailey & Gatrell (p. 165) give a slightly different form for the sample covariogram:

$$\widehat{C}(\boldsymbol{h}) = \frac{1}{n(\boldsymbol{h})} \sum_{\boldsymbol{s_i} - \boldsymbol{s_j} = \boldsymbol{h}} (y_i - \overline{y})(y_j - \overline{y}).$$

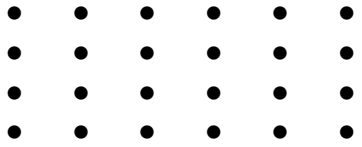
  This form implicitly assumes the mean of the y_i values is the same as the mean of the y_j values, which is approximately true for sufficiently large samples. We will use the first form above as this is what modern software (including **R**) uses.

Example: Consider the following 4x6 lattice as an example.

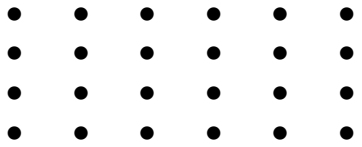
The two covariances $\text{Cov}(Y_i, Y_{i+h})$ and $\text{Cov}(Y_k, Y_{k+h})$ are assumed to be the same.



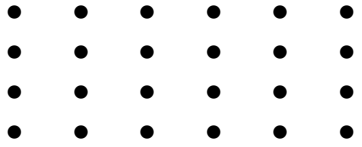
$m_{-(1,0)}$ is the mean of all boxed points
$N_{(1,0)} = 20$



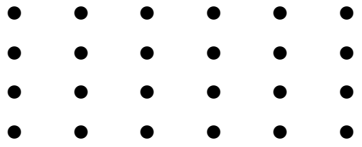
$m_{+(0,1)}$ is the mean of all boxed points
$N_{(0,1)} = 18$



$m_{-(0,1)}$ is the mean of these 18 boxed points



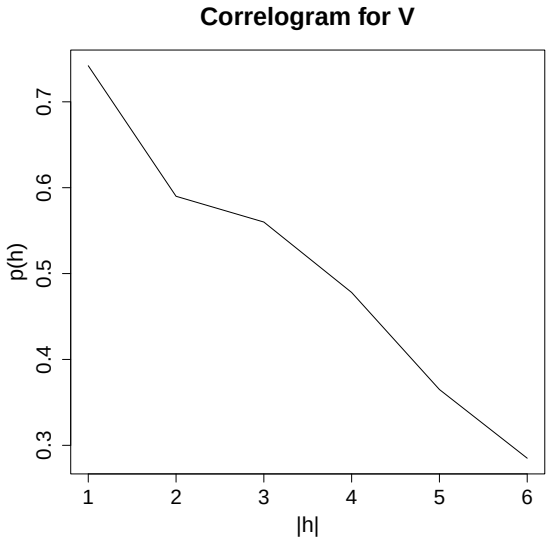
$m_{+(1,1)}$ is the mean of these boxed points
$N_{(1,1)} = 15$



Sample Correlation Function (Correlogram)

The sample correlogram for the Walker Lake 10x10 lattice data and corresponding correlations are given below for the N-S direction. These were found using the **hscatter** function in **R**.

$\boldsymbol{h}$	Correlation Coefficient
(0,1)	0.742
(0,2)	0.590
(0,3)	0.560
(0,4)	0.478
(0,5)	0.365
(0,6)	0.285



Let: $\boldsymbol{s_i} = (s_{1i}, s_{2i}) = i^{th}$ location and

$\boldsymbol{h_{ij}} = \boldsymbol{s_j} - \boldsymbol{s_i}$, the vector connecting points $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$.

The sample correlogram is defined as: $\boxed{\widehat{\rho}(\boldsymbol{h}) = \dfrac{\widehat{C}(\boldsymbol{h})}{\widehat{\sigma}_{-\boldsymbol{h}} \cdot \widehat{\sigma}_{+\boldsymbol{h}}}}$, where:

$$\widehat{\sigma}_{-\boldsymbol{h}} = \sqrt{\frac{1}{n(\boldsymbol{h})} \sum_{i|\boldsymbol{h_{ij}} = \boldsymbol{h}} y_i^2 - m_{-\boldsymbol{h}}^2}$$

$=$ the biased sample standard deviation of all data values oriented at $-\boldsymbol{h}$ from some other data value,

$$\widehat{\sigma}_{+\boldsymbol{h}} = \sqrt{\frac{1}{n(\boldsymbol{h})} \sum_{j|\boldsymbol{h_{ij}} = \boldsymbol{h}} y_j^2 - m_{+\boldsymbol{h}}^2}$$

$=$ the biased sample standard deviation of all data values oriented at $+\boldsymbol{h}$ from some other data value

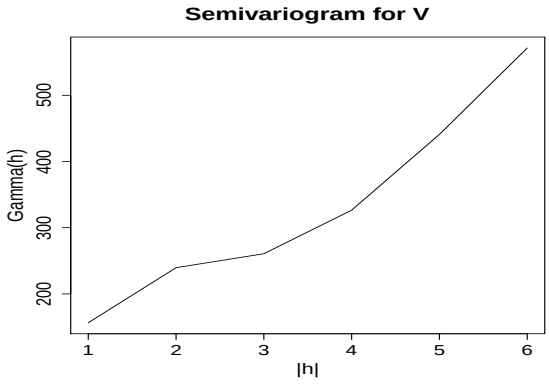
- Again, the reading implies a slightly different form which uses a common standard deviation: $\widehat{\rho}(\boldsymbol{h}) = \widehat{C}(\boldsymbol{h})/\widehat{\sigma}^2$. We will use the form in the boxed expression above.
- Often, $m_{-\boldsymbol{h}}$ & $\widehat{\sigma}_{-\boldsymbol{h}}$ are referred to as the sample mean & standard deviation of the tail values, and $m_{+\boldsymbol{h}}$ & $\widehat{\sigma}_{+\boldsymbol{h}}$ are referred to as the sample mean & standard deviation of the head values.

Variogram (& Semivariogram)

The variogram is the variance of the *difference* between random variables at two units (locations), given by: $\text{Var}(Y_i - Y_j)$ for two sites $\boldsymbol{s_i}$ & $\boldsymbol{s_j}$. The semivariogram is one half the variogram. Both are just another measure of spatial autocorrelation. If two units are close together, their difference will typically be small, as would the variance of the difference. As units get further apart, their differences get larger and usually the variance of the difference gets larger.

Again, for the Walker Lake data in the N-S direction, the semivariogram plot and a table of the corresponding values is given below:

$\boldsymbol{h}$	Semivariogram (ppm^2)
(0,1)	156.4
(0,2)	239.6
(0,3)	260.7
(0,4)	326.5
(0,5)	440.9
(0,6)	571.4



This curve is the exact mirror image of the covariance function given earlier. At lag zero, the semivariogram is zero because $Y_i - Y_i = 0$; that is, the value at one spot is a constant and has no variance. [Note that this assumes no replication at that spot!] The curve may reach a plateau, indicating that past a certain distance, the correlation between two units is zero. If it does reach a plateau, the value of the semivariogram at that point is just the variance of Y. Why?

Let: $\boldsymbol{s_i} = (s_{1i}, s_{2i}) = i^{th}$ location and

$\boldsymbol{h_{ij}} = \boldsymbol{s_j} - \boldsymbol{s_i}$, the vector connecting points $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$.

The sample semivariogram is defined as: $\boxed{\widehat{\gamma}(h) = \frac{1}{2n(\boldsymbol{h})} \sum_{(i,j)|\boldsymbol{h_{ij}} = \boldsymbol{h}} (y_i - y_j)^2}$.

Relationship Between Covariance Function and Semivariogram

Recall that the theoretical covariance function and semivariogram are defined as:

$$\textbf{Covariance Function}: \quad C_{ij} = \mathrm{Cov}(Y_i, Y_j).$$

$$\textbf{Semivariogram}: \quad \gamma_{ij} = \frac{1}{2}\mathrm{Var}(Y_i - Y_j)$$

$$=$$

$$=$$

Assuming 2nd-order stationarity (what does this mean?), the relationship between the semivariogram and the covariance function is given below:

$$\gamma(\boldsymbol{h}) = \frac{1}{2}\mathrm{Var}(Y_i - Y_j) = \frac{1}{2}\{\mathrm{Var}(Y_i) + \mathrm{Var}(Y_j) - 2\mathrm{Cov}(Y_i, Y_j)\}.$$

If $\mathrm{Var}(Y_i) = \mathrm{Var}(Y_j) = \mathrm{Var}(Y) = \sigma^2$ and $\mathrm{C}(Y_i, Y_j) = \mathrm{C}(\boldsymbol{h})$ for all i, j such that $\boldsymbol{h_{ij}} = \boldsymbol{h}$, then:

$$\boxed{\gamma(\boldsymbol{h})} = \frac{1}{2}(2\sigma^2 - 2C(\boldsymbol{h})) = \boxed{\sigma^2 - C(\boldsymbol{h})}.$$

- Therefore, at $\boldsymbol{h} = \boldsymbol{0}$, the semivariogram is: $\gamma(\boldsymbol{0}) = \sigma^2 - C(\boldsymbol{0}) = \sigma^2 - \sigma^2 = 0$.
- For $\boldsymbol{h}$ large, where $\mathrm{C}(\boldsymbol{h}) = 0$, the semivariogram is: $\gamma(\boldsymbol{h}) = \sigma^2 - C(\boldsymbol{h}) = \sigma^2$.

Comparison of Covariance Function & Semivariogram

- From a practical standpoint, the semivariogram is generally preferred as a measure of spatial correlation because it tends to be "smoother" than the covariance function.
- The semivariogram also has the computational advantage that only pairs of values h units apart need to be considered; we don't need the overall means, $m_{+\boldsymbol{h}}$ and $m_{-\boldsymbol{h}}$. So if there is some trend of which we are unaware, the semivariogram is not affected as much as the correlation function would be.
- For purposes of interpretation and discussion, it may be more natural to use correlations and the correlogram, as this is the common way in statistics of viewing association.

Cross-Covariance Function, Correlogram, Variogram

For a single response variable Y, we have seen that h-scatterplots provide a graphical description of spatial autocorrelation, and that covariograms, correlograms, and variograms provide numerical descriptions of spatial autocorrelation. It is often of interest to characterize the degree of spatial correlation across the responses of *two* variables, not just one. In this bivariate setting, analogous descriptions to those defined above, namely cross h-scatterplots, cross-covariograms, cross-correlograms, and cross-variograms, are used. This handout defines each of these "cross-functions" and illustrates their use with a small hypothetical example and with the V & U Walker Lake concentration data.

Cross h-Scatterplots: As with h-scatterplots, cross h-scatterplots are a visual tool for examining the correlation between a response variable at a given point and a response variable at neighboring points. The difference is that the pair of responses are from two different variables.

- Let:
$$\begin{aligned} \boldsymbol{s_i} &= (s_{1i}, s_{2i}) = i^{th} \text{ location}, \\ \boldsymbol{h_{ij}} &= \boldsymbol{s_j} - \boldsymbol{s_i}, \text{ the vector connecting points } \boldsymbol{s_i} \text{ and } \boldsymbol{s_j}, \\ U_i &= \text{the response of variable } U \text{ at location } i, \\ V_i &= \text{the response of variable } V \text{ at location } i,\ i = 1, \ldots, n. \end{aligned}$$

- The cross $\boldsymbol{h}$-scatterplot plots U_j versus V_i for all pairs of locations $\boldsymbol{s_i}$ & $\boldsymbol{s_j}$ separated in distance and direction by the vector $\boldsymbol{h}$.

Example: Suppose we are interested in two variables U & V on a 3x3 spatial grid. These variables might be concentrations of two chemicals, say iron and magnesium, in the soil. It may be of interest to know how the concentrations of these two chemicals vary with respect to each other spatially. Below is a 3x3 table representing the 9 Fe and Mg values collected over the 9 sites. The sites are labeled from 1 to 9 arbitrarily.

- If we want the cross correlation at neighboring sites in the N-S direction, we would use a cross $\boldsymbol{h}$ = (0,1)-scatterplot.

  The resulting pairs are given by:

U_1	U_2	U_3	Fe
V_1	V_2	V_3	Mg
U_4	U_5	U_6	
V_4	V_5	V_6	
U_7	U_8	U_9	
V_7	V_8	V_9	

- If we want the cross correlation at neighboring sites in the N-S direction, separated by 2 units, we use a cross $\boldsymbol{h}$ = (0,2)-scatterplot.

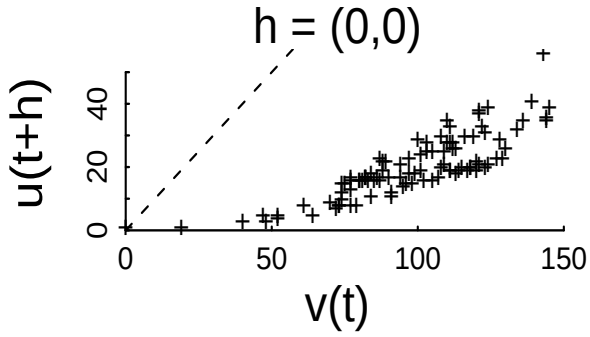
  The resulting pairs are given by:

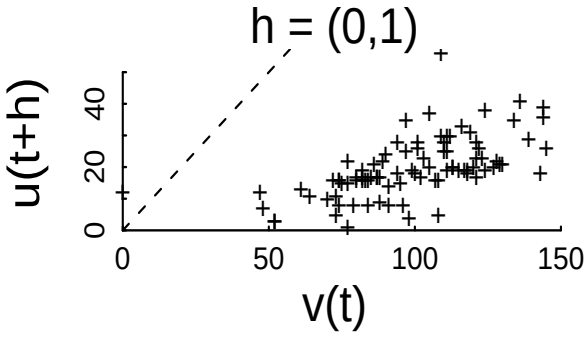
U_1	U_2	U_3	Fe
V_1	V_2	V_3	Mg
U_4	U_5	U_6	
V_4	V_5	V_6	
U_7	U_8	U_9	
V_7	V_8	V_9	

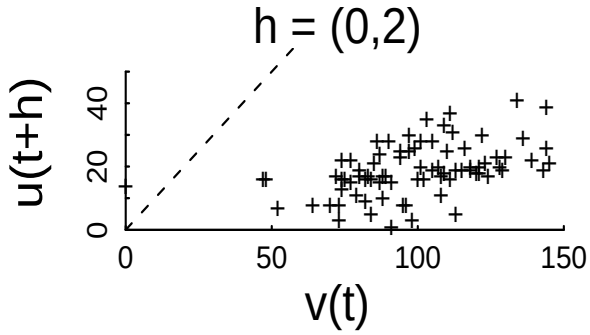
- What does a cross $\boldsymbol{h} = (0,0)$-scatterplot mean?

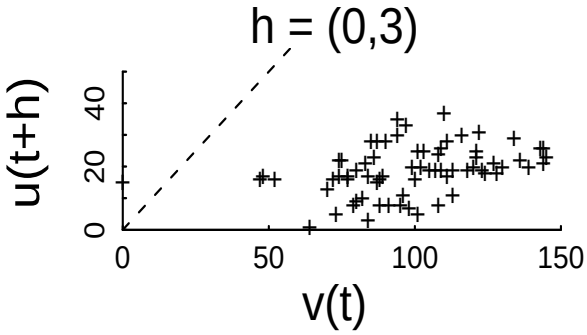
- As with h-scatterplots, cross h-scatterplots can be done in any direction at any distance.

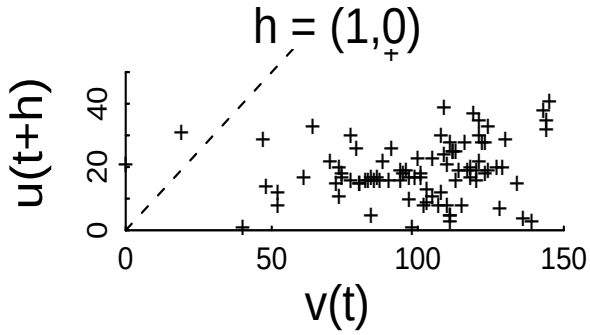
Example: Walker Lake Data Cross h-Scatterplots: Using the 10x10 illustrative data set of U & V concentration values, the h-scatterplots below were created to examine the relationship between U and V in both the north-south and east-west directions. What do you notice?











The **R** function **hscatter** was used to generate this set of plots, and the code is given at the end of this handout.

Cross-Covariance, Cross-Correlation, Cross-Variogram: These are the analogous "cross" measures of spatial correlation between two variables. To illustrate how these various measures are computed, reconsider the example given earlier on iron (Fe) and magnesium (Mg) concentrations measured on a 3x3 grid. The data for both variables are given below in the 3x3 table, with the boxed values representing the Fe measurements.

[7]	[8]	[8]	Fe (U)
3	4	5	Mg (V)
[9]	[10]	[10]	
5	4	6	
[10]	[12]	[12]	
5	6	5	

Cross-Covariance Function: The cross-covariance function between variables U & V is defined to be: $C_{UV}(\boldsymbol{h}) = E[(U_i - E(U_i))(V_j - E(V_j))]$, where $\boldsymbol{s_j - s_i = h}$. This function is estimated by the sample cross-covariance function, given by:

$$\widehat{C}_{UV}(\boldsymbol{h}) = \frac{1}{n(\boldsymbol{h})} \sum_{(i,j)|\boldsymbol{h_{ij}} = \boldsymbol{h}} u_i \cdot v_j - m_{u,-\boldsymbol{h}} \cdot m_{v,+\boldsymbol{h}}, \text{ where:}$$

$$\begin{aligned} n(\boldsymbol{h}) &= \text{the number of pairs } \boldsymbol{h} \text{ units apart} \\ m_{v,+\boldsymbol{h}} &= \text{the mean of all } v_j \text{ values in the sum} \\ m_{u,-\boldsymbol{h}} &= \text{the mean of all } u_i \text{ values in the sum} \end{aligned}$$

- Notation: We will use $\widehat{C}_{UV}(\boldsymbol{h})$ to represent the sample (empirical) cross-covariance function, reserving $C_{UV}(\boldsymbol{h})$ as the population measure of cross-covariance.
- This function is not necessarily symmetric. In other words, it is not necessarily the case that $\widehat{C}_{UV}(\boldsymbol{h}) = \widehat{C}_{VU}(\boldsymbol{h})$.

To demonstrate computation of the cross-covariance function, consider the iron-magnesium data given above. Suppose we want to calculate the cross-covariance in the (1,0)-direction. For this example, there are 6 pairs of (U,V) values as shown in the data. Computing:

$$\begin{aligned} n(1,0) &= 6 \\ m_{u,-(1,0)} &= \frac{1}{6}(7+8+9+10+10+12) = 9.\overline{3}, \\ m_{v,+(1,0)} &= \frac{1}{6}(4+5+4+6+6+5) = 5, \\ \widehat{C}_{UV}(1,0) &= \frac{1}{6}(28+40+36+60+60+60) - (9.\overline{3})(5) \\ &= \frac{2}{3}. \end{aligned}$$

[7]	[8]	[8]	Fe (U)
3	4	5	Mg (V)
[9]	[10]	[10]	
5	4	6	
[10]	[12]	[12]	
5	6	5	

What does this number mean?

Cross-Correlation Function: The cross-correlation function between variables U & V is defined to be: $\rho_{UV}(\boldsymbol{h}) = C_{UV}(\boldsymbol{h})/\sigma_{U,-\boldsymbol{h}}\sigma_{V,+\boldsymbol{h}}$, where: $\boldsymbol{s_j - s_i = h}$. This function is estimated by the sample cross-correlation function, given by:

$$\boxed{\widehat{\rho}_{UV}(\boldsymbol{h}) = \frac{\widehat{C}_{UV}(\boldsymbol{h})}{\widehat{\sigma}_{U,-\boldsymbol{h}}\widehat{\sigma}_{V,+\boldsymbol{h}}}}, \text{ where:}$$

$$\begin{aligned} \widehat{\sigma}_{v,+\boldsymbol{h}} &= \text{the standard deviation of all } v_j \text{ values in the sum,} \\ \widehat{\sigma}_{u,-\boldsymbol{h}} &= \text{the standard deviation of all } u_i \text{ values in the sum.} \end{aligned}$$

Back to the Example: Computing the cross-correlation function for $\boldsymbol{h}$ = (1,0):

$$\begin{aligned} \widehat{\sigma}^2_{U,-(1,0)} &= \frac{1}{6}(7^2 + 8^2 + 9^2 + 10^2 + 10^2 + 12^2) - (9.\overline{3})^2 = 2.\overline{5} \\ \widehat{\sigma}^2_{V,+(1,0)} &= \frac{1}{6}(4^2 + 5^2 + 4^2 + 6^2 + 6^2 + 5^2) - (5)^2 = 0.\overline{6} \\ \widehat{\rho}_{UV}(1,0) &= \frac{2/3}{\sqrt{(2.\overline{5})(2/3)}} = \underline{0.51}. \end{aligned}$$

- This value is just the sample correlation coefficient between U_i and V_j where U values are included only in the tail positions and V values are included only in the head positions.
- As with the cross-covariance function, this cross-correlation function is not symmetric (i.e.: if we were to calculate $\widehat{\rho}_{VU}(1,0)$, it would likely differ from $\widehat{\rho}_{UV}(1,0)$).

Cross-Variogram: The cross-variogram can be defined in two different ways, although the texts consider only one of these. Both of these definitions are generalizations of the variogram, defined earlier as: $2\gamma_{VV}(\boldsymbol{h}) = \text{Var}(V_i - V_j)$, where: $\boldsymbol{h_{ij} = h}$.

1. $2\gamma_{1UV}(\boldsymbol{h}) = \text{Cov}(U_i - U_j, V_i - V_j)$, where: $\boldsymbol{h_{ij} = h}$.
2. $2\gamma_{2UV}(\boldsymbol{h}) = \text{Var}(U_i - V_j)$, where: $\boldsymbol{h_{ij} = h}$.

These are both population cross-variograms. Both Bailey & Gatrell and Isaaks & Srivastava define the cross-variogram as the first form given above (see section 6.2.2, bottom of page 213, for B & G; see p. 62 for I & S). Using this definition, the sample cross-semivariogram is defined as:

$$\widehat{\gamma}_{1UV}(\boldsymbol{h}) = \frac{1}{2n(\boldsymbol{h})} \sum_{(i,j)|\boldsymbol{h_{ij} = h}} (u_i - u_j)(v_i - v_j).$$

Under this definition, $\widehat{\gamma}_{1UV}(\boldsymbol{h})$ is one-half the sample covariance of the differences in the two variables. Computing this for the Fe-Mg example (in the (1,0) direction):

$$\begin{aligned} \widehat{\gamma}_{1UV}(\boldsymbol{h}) &= \frac{1}{2(6)}[(7-8)(3-4) + (8-8)(4-5) + (9-10)(5-4) + (10-10)(4-6) \\ &\quad +(10-12)(5-6) + (12-12)(6-5)] = \frac{1}{12}(1 - 1 + 2) = \underline{\frac{1}{6}}. \end{aligned}$$

- What does this mean? It is interpreted as the covariance between the change in iron and the change in magnesium as you move one unit to the right (east). In other words, it addresses whether the changes in the two variables are spatially similar.
- Note that by definition, this cross-semivariogram is symmetric (i.e.: $\widehat{\gamma}_{1UV}(\boldsymbol{h}) = \widehat{\gamma}_{1VU}(\boldsymbol{h})$). Why?

Comparison with Semivariograms: The semivariograms for the Fe (U) and Mg (V) concentration variables are given by:

$\widehat{\gamma}_U(1,0) \;=$

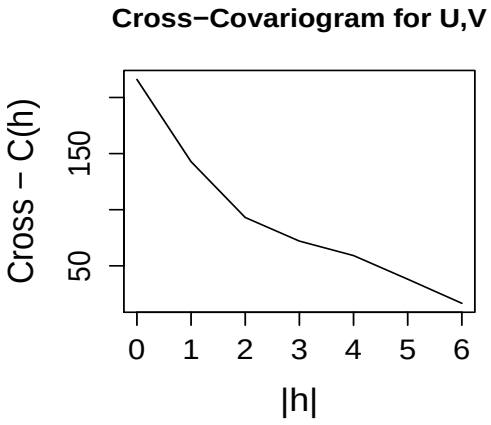
$\widehat{\gamma}_V(1,0) \;=$

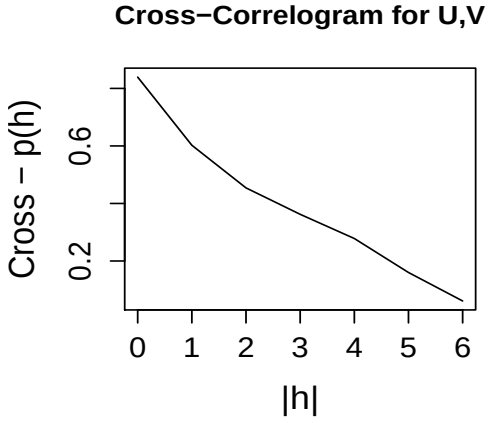
Thus, this covariance between the change in iron and the change in magnesium is small relative to the variances of the change in iron and the change in magnesium.

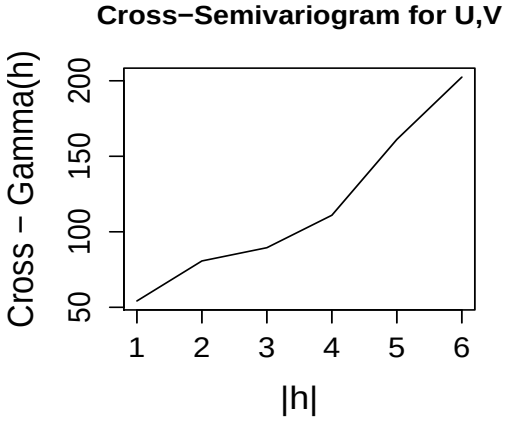
- The second definition of cross-variograms will not be considered here.

Cross-Functions for the Walker Lake Data: A table of cross-covariances, cross-correlations, and cross-variograms and corresponding plots of these values versus their lags are shown below. These plots indicate the association between U & V in the N-S direction.

$\boldsymbol{h}$	Cross Correlation Coefficient	Cross Covariance ppm^2	Cross Variogram ppm^2
(0,0)	0.84	216.1	0.0
(0,1)	0.60	142.8	54.2
(0,2)	0.45	93.1	80.7
(0,3)	0.36	72.0	89.5
(0,4)	0.28	59.1	111.0
(0,5)	0.16	38.1	161.2
(0,6)	0.06	16.6	202.4







R Code Used to Generate Cross h-Scatterplots and Cross Functions

```
# Plots the cross h-scatterplots for the Walker Lake (n=100) V&U-data, and
# plots the cross-covariogram, cross-correlogram, and cross-semivariogram
# functions for the U,V data.  The hscatter function was used to generate
# the cross h-scatterplots.
# ========================================================================
walk100 <- read.table("Data/walk100.txt",header=T) # Reads in walk100 Walker Lake data.
x <- walk100$x                          # x is set to the x-values of "walk100".
y <- 11 - walk100$y                     # y is set to 11 - the y-values of "walk100".
v <- walk100$v                          # v is set to the V-values of "walk100".
u <- walk100$u                          # u is set to the U-values of "walk100".

# Creates the 5 cross-h scatterplots on page 39 of the class notes
# ================================================================
par(mfrow=c(3,2))                       # Sets up a 3x2 graphics window.
hscatter(x,y,v,u,c(0,0))                # Produces a scatterplot of u vs. v.
hscatter(x,y,v,u,c(0,1))                # Produces a cross h=(0,1)-scatterplot of u vs. v.
hscatter(x,y,v,u,c(0,2))                # Produces a cross h=(0,2)-scatterplot of u vs. v.
hscatter(x,y,v,u,c(0,3))                # Produces a cross h=(0,3)-scatterplot of u vs. v.
hscatter(x,y,v,u,c(1,0))                # Produces a cross h=(1,0)-scatterplot of u vs. v.

# Plots the 3 cross-functions on page 42 of the class notes
# ========================================================
h <- c(0,1,2,3,4,5,6)                                  # Sets a vector of h-values.
out <- matrix(nrow=length(h),ncol=3)                   # Defines an hx3 blank matrix.
for (i in 0:6) out[i+1,] <- as.numeric(                # Loops through distances of 0 to 6 and
  hscatter(x,y,v,u,c(0,i)))                            #   calculates cross-functions for each.
cch <- out[,1]; cph <- out[,2]; cgh <- out[,3]         # Defines the vectors of cross-covariances,
                                                       #   cross-corr's, & cross-semivariograms.
par(mfrow=c(2,2))                                      # Sets up a 2x2 graphics window.
plot(h,cch,type="n",axes=F,xlab="",ylab="")            # Creates a completely blank plot.
plot(h,cch,type="l",xlab="|h|",ylab="Cross - C(h)",# Plots the cross-correlations vs. h
  cex.lab=1.5,cex.axis=1.3,cex.main=1,                 #   with sizes controlled by "cex"
  main="Cross-Covariogram for U,V")                    #   and a title.
plot(h,cph,type="l",xlab="|h|",ylab="Cross - p(h)",# Plots the cross-correlogram vs. h
  cex.lab=1.5,cex.axis=1.3,cex.main=1,                 #   with sizes controlled by "cex"
  main="Cross-Correlogram for U,V")                    #   and a title.
plot(h[-1],cgh[-1],type="l",xlab="|h|",cex.lab=1.5,# Plots the cross-semivariogram vs. h
  ylab="Cross - Gamma(h)",cex.axis=1.3,cex.main=       #   with sizes controlled by "cex".
  1,main="Cross-Semivariogram for U,V")                # Puts a title on the plot.
```

Notes on Walker Lake Data Sets (Ch. 5,6 - I & S)

The exhaustive (78000 sites) and sample (470 sites) Walker Lake data sets are explored in Chapters 5 & 6 of Isaaks & Srivastava respectively. Recall that for each of these, there are 2 concentration variables labeled U & V. Additionally, there is an indicator variable T for two habitat types (T=1 or 2). This handout lists some of the key features of these data sets.

Exhaustive Data Set

- Both the V and U data are highly right skewed, having many 0 concentrations (Figures 5.1, 5.3 below, Tables 5.1, 5.2). When broken up by the two types in T, the V & U distributions are still right skewed, but much less so for type 2 values (Figures 5.6, 5.7).

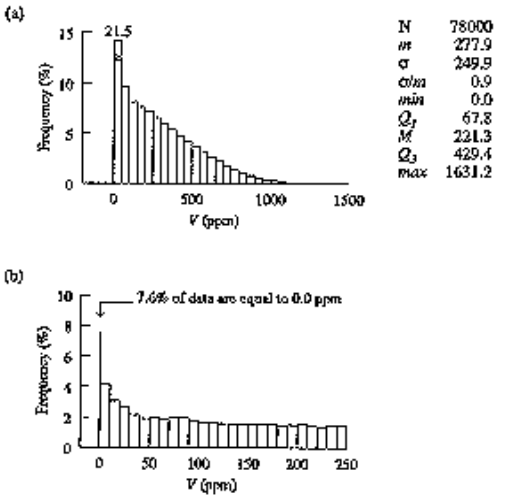


Figure 5.1 Histogram and univariate statistics of the 78,000 V values using a class width of 50 ppm in (a) and a class width of 10 ppm in (b).

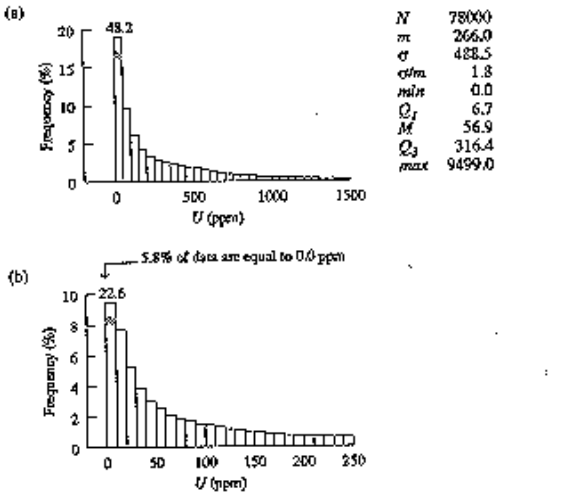


Figure 5.3 Histogram and univariate statistics of the 78,000 U values using a class width of 50 ppm in (a) and a class width of 10 ppm in (b).

- Normal quantile plots indicate a lack of normality for both V & U (Figures 5.2, 5.4). A log transform (common with concentration data) attempted to straighten out the pattern in the normal quantile plot, but seemed to make matters worse. Note that the log transform (shown below) reversed the direction of the curvature in the QQ-plot, suggesting that a log transform may have been too severe. Can you suggest some other less severe transformations which might have normalized the data?

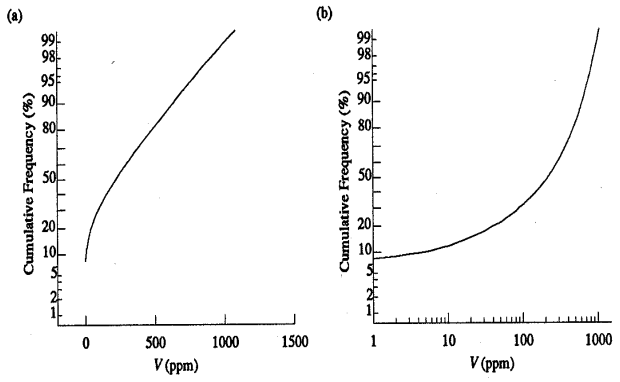


Figure 5.2 The normal probability plot of the 78,000 V values is given in (a), and the lognormal probability plot in (b).

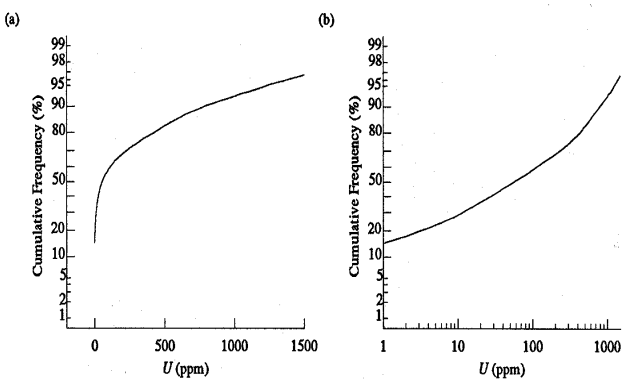


Figure 5.4 The normal probability plot of the 78,000 U values is given in (a), and the lognormal probability plot in (b).

- The distribution of T is best given in an indicator map, as shown in Figure 5.5 to the right, and on page 74. The text point out that most of the Type 1 areas are comprised of Walker Lake itself and a valley running SE away from the lake. Histograms, normal quantile plots, and other continuous data summaries are not appropriate for binary data such as T.

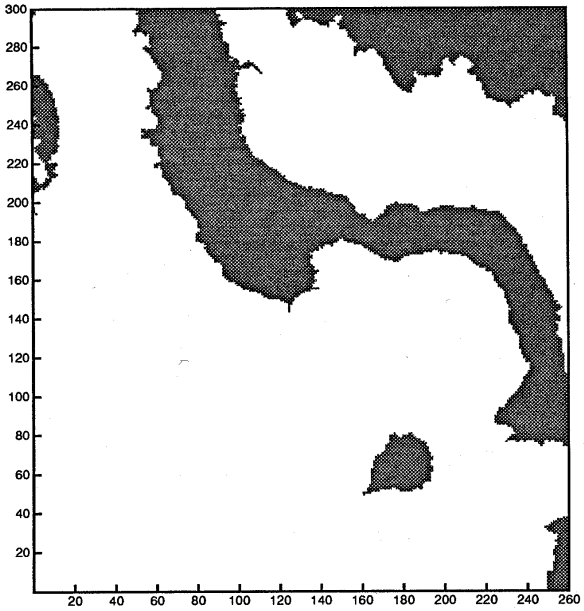


Figure 5.5 Location map of the third variable, T. The cross hatched area shows the location of the type 1 T values. The remaining area corresponds to type 2 values of T.

- Note the difficulties with constructing contour plots and histograms with large data sets. Note especially the problems involved in displaying data which have lots of zeros - this is a frequent occurence with concentration data as well as percentage data.

- The variables V & U are moderately positively associated, although not linearly. It is clear from the scatterplots of V vs. U (Figure 5.8) that the values of V are larger for T=2 than for T=1. Hence T may be an important covariate in understanding the relationship between V & U.

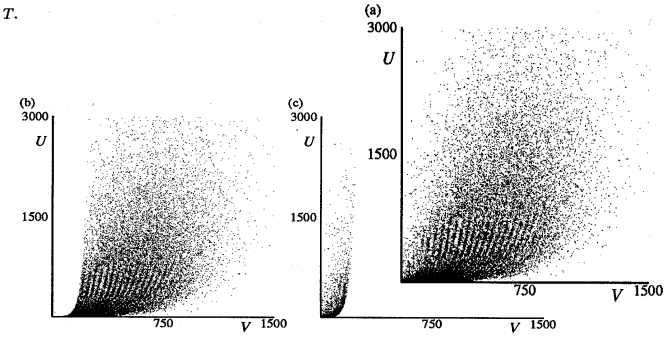


Figure 5.8 (a) Scatterplot of all 78,000 values of V and U (b) 60,384 values of V and U for $T = 2$, and (c) 17,616 values of V and U for $T = 1$.

- The series of indicator maps (Figures 5.9, 5.10 shown below) for V & U indicate that the smaller values for V & U occur around Walker Lake and the valley next to it (i.e.: Type I data). The large U values are somewhat more scattered (less spatial continuity) than the large V values, although both display definite and similar spatial patterns.

Figure 5.9e Indicator map of V for the fifth decile, 221.25 ppm.

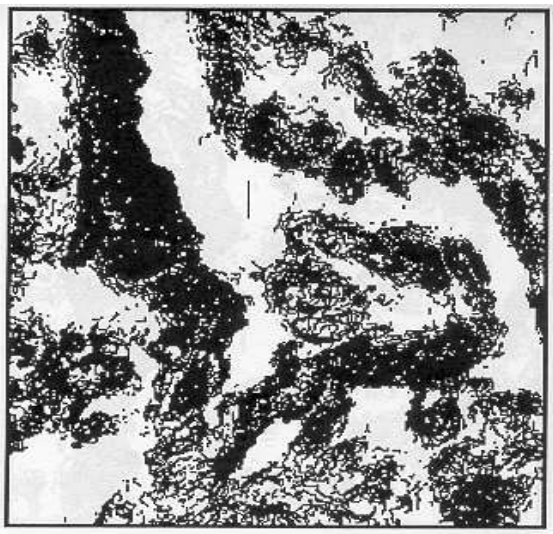


Figure 5.10e Indicator map of U for the fifth decile, 56.90 ppm.

- Using first 10x10 and then 20x20 nonoverlapping moving windows, means & variances for the 780 windows (100 data per window) and 195 windows (400 data per window) were calculated for the V data and U data. Contour maps for the resulting means & variances were drawn for both V & U (Figures 5.11 - 5.16). These maps show clear designations between the lake and an area of high values immediately west of the lake. Plots of the moving window means vs. the standard deviations (given to the right) show for both V & U that some relationship exists between the mean and variance (Figure 5.13). The relationship is stronger for the U data where the mean and standard deviation are roughly proportional, but is clearly present for the V data as well (see page 93). These relationships were not studied for the two T types separately, although they probably should be.

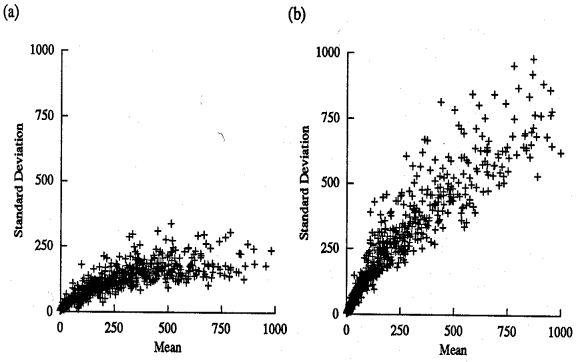


Figure 5.13 Scatterplot of standard deviations versus means. In (a), the standard deviations of V within 10 x 10 m^2 blocks are plotted versus the mean of V within the same blocks. The corresponding plot for the U values is shown in (b).

- Because there are so many distances and directions of spatial continuity which could be considered, the values for the covariance function at all distances and directions corresponding to nodes (1 meter apart) of a 200x200m grid centered at $\boldsymbol{h} = (0,0)$ were computed. In other words, 40000 C($\boldsymbol{h}$) values were computed (for different $\boldsymbol{h}$ vectors) and then contoured to examine the degree of spatial correlation at various distances and directions (see Figures 5.17 & 5.19 for the contour plots and Figures 5.18 & 5.20 for covariograms in selected directions). For both V & U, the direction of maximum change (minimum spatial correlation) was approximately 14° north of east, and the direction of minimum change was approximately 14° west of north. The plots for U, where the directional effects were greater, are given below.

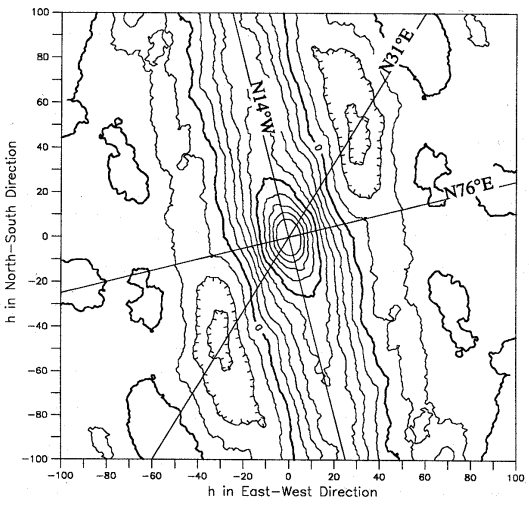


Figure 5.19 Contour map of the exhaustive covariance function for U. The values contoured are the covariance values of all possible h-scatterplots in every direction to a distance of at least 100 meters. The contour interval is 10,000 ppm^2. The lines N14°W and N76°E are the directions of maximum and minimum continuity; the line N31°E lies along a direction midway between these two axes. Profiles of the covariance function along these directions are shown in Figure 5.20.

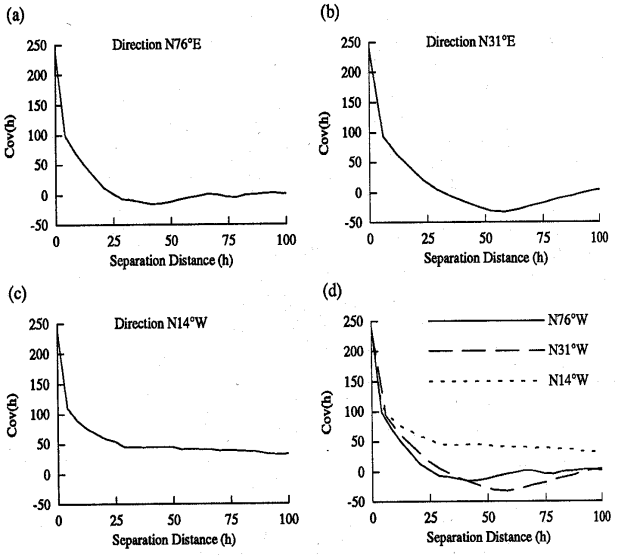


Figure 5.20 (a), (b), and (c) are cross sectional profiles of the exhaustive covariance function for U in three specific directions. All three profiles are plotted in (d). The vertical axis on all the plots is labeled in units of thousands of ppm^2.

- Computation of the cross-covariance function contours gave similar directions of maximum and minimum cross-covariances (Figures 5.21, 5.22).
- When the covariances and cross-covariances were recalculated for the two types of data separately, the directions of maximum and minimum correlation were again similar for U, V, and UV; however, the magnitude of the correlations in these directions were markedly different. The difference in the strengths of the correlations in the directions of minimum and maximum correlation was much more pronounced for the Type 1 data (Figures 5.23, 5.24).

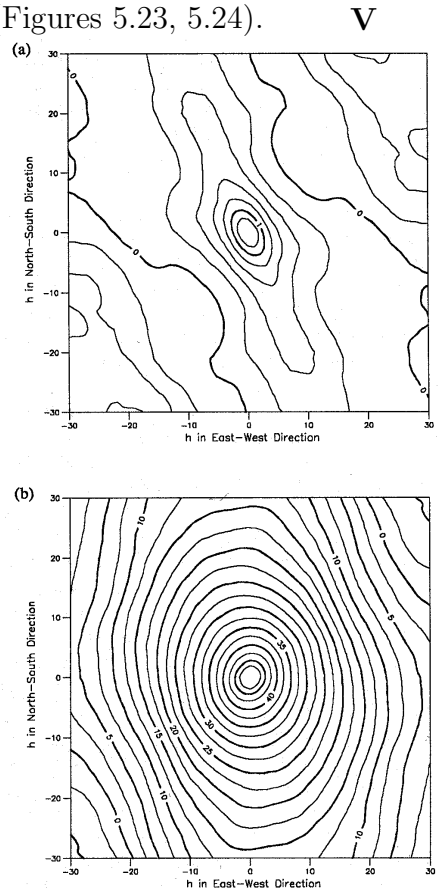


Figure 5.23 Contour maps of the exhaustive covariance functions of V for each T type. (a) shows the contoured covariance function of V for type 1, and (b) for type 2. The contour labels are in thousands of ppm^2.

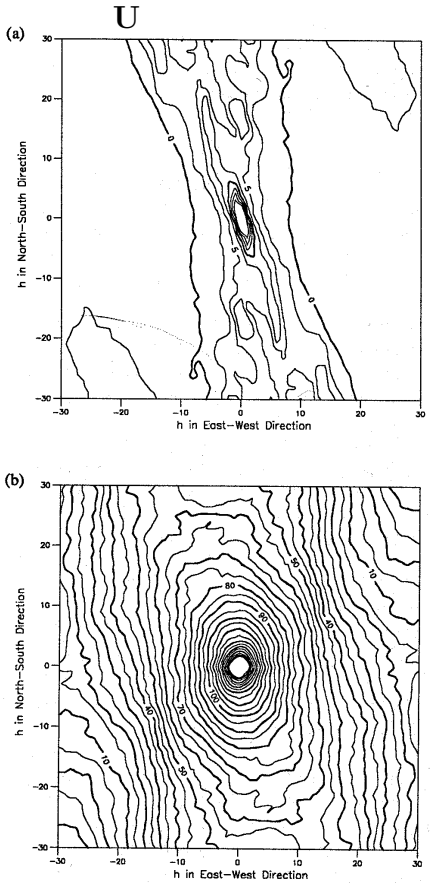
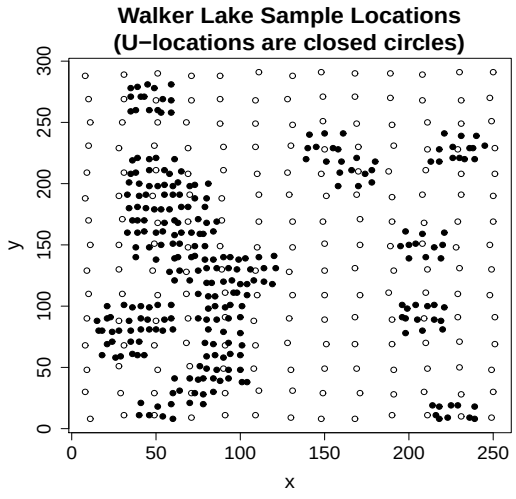


Figure 5.24 Contour maps of the exhaustive covariance functions of U for each type of T. (a) shows the contoured covariance function of U for type 1, and (b) for type 2. The contour labels are in thousands of ppm^2.

Sample Data Set

- The most striking difference between the summary statistics for U & V from the sample data and exhaustive data is that the values from the sample data are much larger. This was due to the sampling plan used with the sample data set, where the goal was to purposely oversample areas with larger concentrations. The difference is especially large for the U values. [See Tables 6.2 & 6.3 to compare the two data sets.]



- The U distribution values were more skewed to the right than the V values (Figures 6.3, 6.5). This may be an artifact of sampling only large U values in the 2nd and 3rd stages of the sampling.

Table 6.2 Comparison of V stat's by sampling program

	Exhaustive	Sample Statistics		
	Statistics	Program 1	Program 2	Program 3
n	78,000	195	150	125
m ($\overline{x}$)	278	275	502	610
s	250	250	295	247
CV	0.90	0.91	0.59	0.41
min	0	0	0	0
Q_1	68	62	269	440
M	221	209	518	608
Q_3	429	426	675	781
max	1,631	975	1,528	1,392

Table 6.3 Comparison of U stat's by sampling program

	Exhaustive	Sample Statistics		
	Statistics	Program 1	Program 2	Program 3
n	78,000		150	125
m ($\overline{x}$)	266		601	628
s	489		801	724
CV	1.84	Not	1.33	1.15
min	0	Available	0	0
Q_1	7		67	111
M	57		254	397
Q_3	316		782	936
max	9,500		3,739	5,190

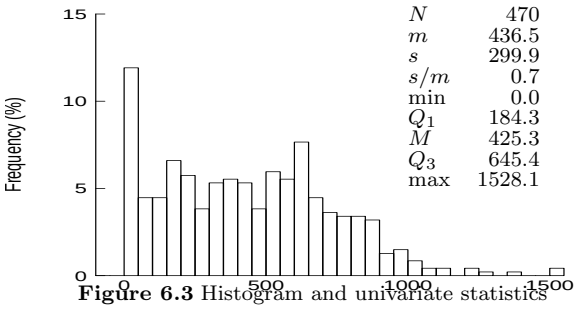


Figure 6.3 Histogram and univariate statistics of the 470 V sample values

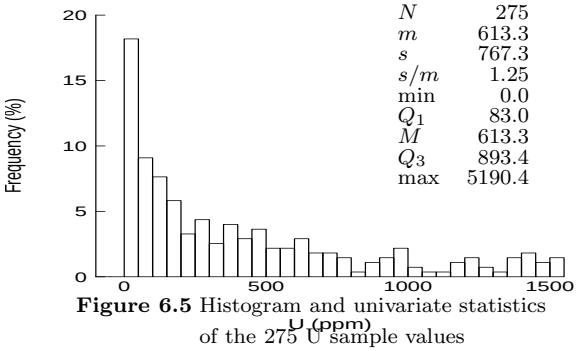


Figure 6.5 Histogram and univariate statistics of the 275 U sample values

- The U & V values are significantly larger for the Type 2 data than for the Type 1 data (Table 6.4). This was true for the exhaustive data set as well.

- There is a moderately positive linear relationship between U & V ($r = 0.55$) (Figure 6.7). This might indicate that the sampling plan to find large values of U was a good one.

Table 6.4 Comparison of V and U statistics by sample type

	V		U	
	Type 1	Type 2	Type 1	Type 2
n	45	425	4	271
m ($\overline{x}$)	40	479	429	616
s	52	284	479	772
CV	1.29	0.59	1.12	1.25
min	0	0	0	0
Q_1	0	241	15	85
M	18	477	367	335
Q_3	72	663	906	893
max	195	1,528	983	5,190

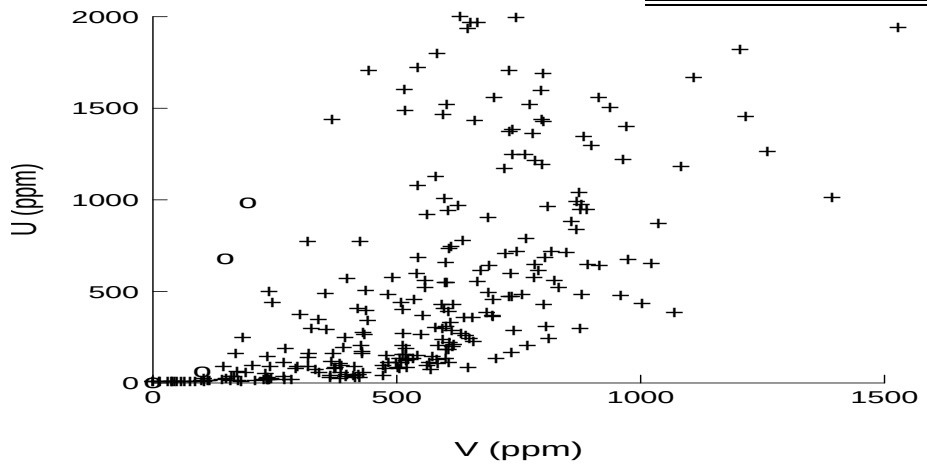


Figure 6.7: A scatterplot of the 275 U and V sample data. The type 1 points are shown with the symbol "o" while the type 2 points are shown with a "+".

- The contour maps for U & V at the end of Chapter 6 highlight the fact that most of the largest U & V values occurred in the Wassuk range running N-S on the west side of the region (Figures 6.9, 6.11 to the right). In fact, there seemed to be three pockets of particularly high values in that range, especially for the U values.

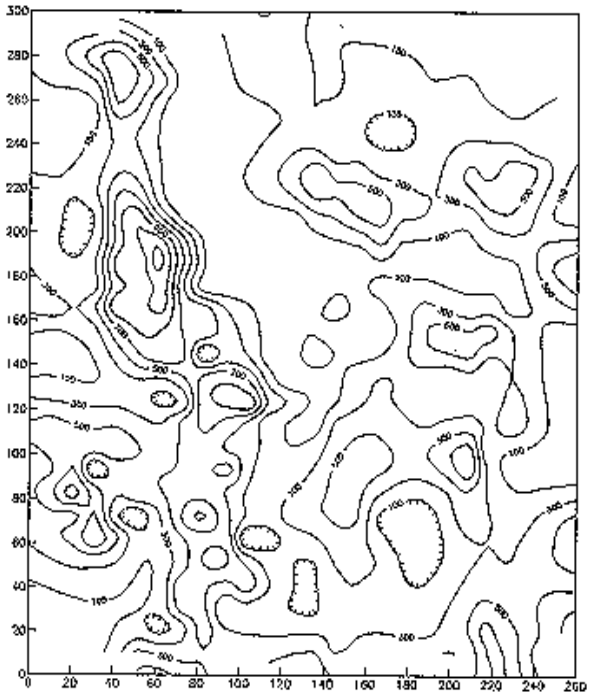


Figure 6.9 Contour map of the 470 V sample data. The contour interval is 200 ppm and begins at 100 ppm.

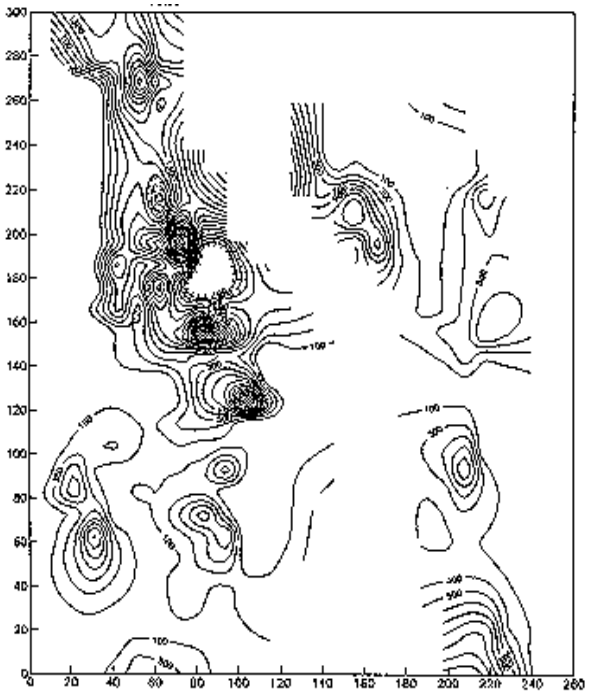


Figure 6.11 Contour map of the 275 U sample data. The contour interval is 200 ppm and begins at 100 ppm.

- Finally, there is a strong proportional effect for both U & V. The moving window means and standard deviations were highly correlated ($r = 0.81, 0.88$) for V & U respectively (Figures 6.12 - 6.14). The roughly proportional relationship between the mean and standard deviation suggests what type of transformation?

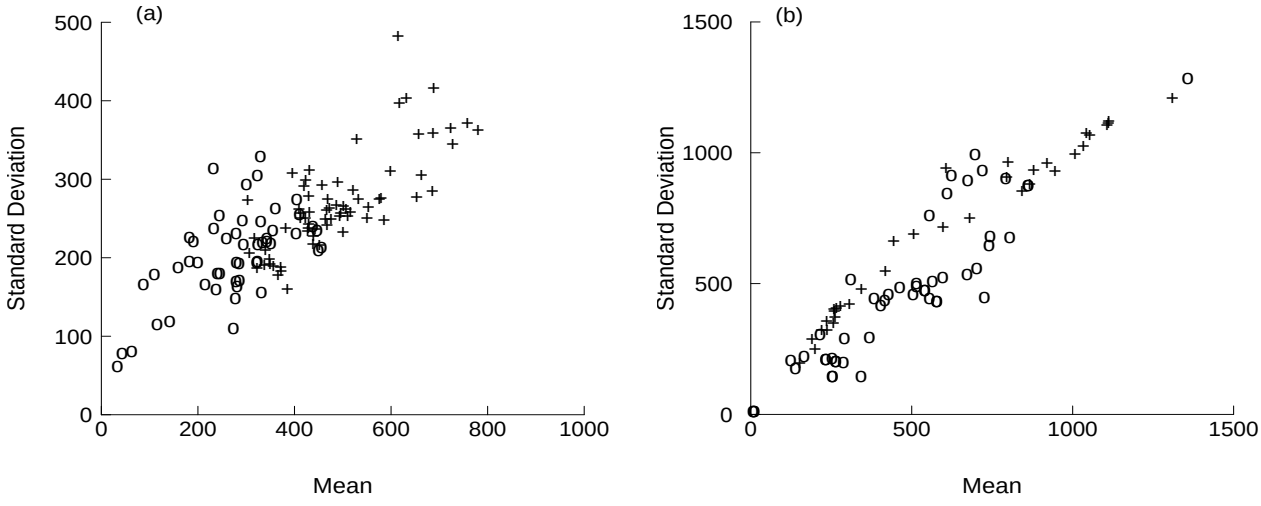
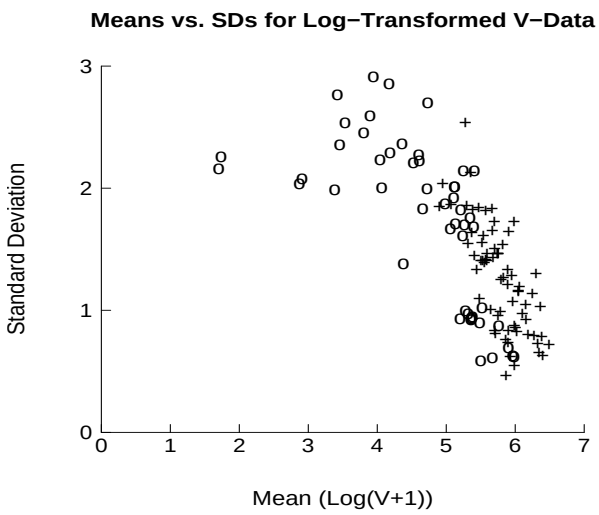
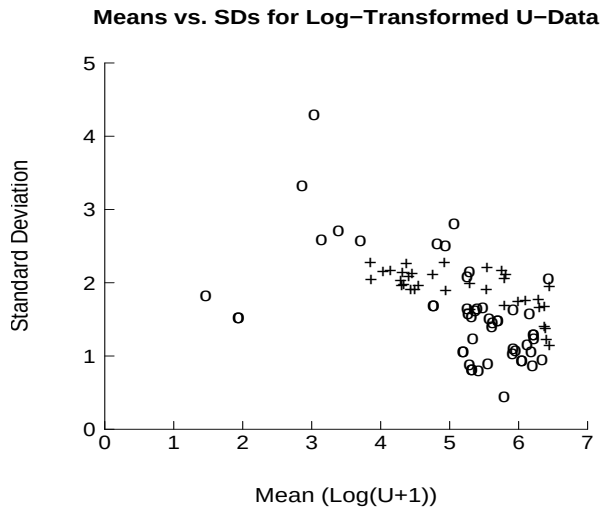


Figure 6.13: Scatterplots of moving window standard deviations versus means for (a) the 470 sample V data and (b) the 275 U sample data. For moving windows containing less than 20 data values the standard deviation and mean were plotted using the "o" symbol, otherwise a "+" symbol was used. Note that the scale of the two plots differs.

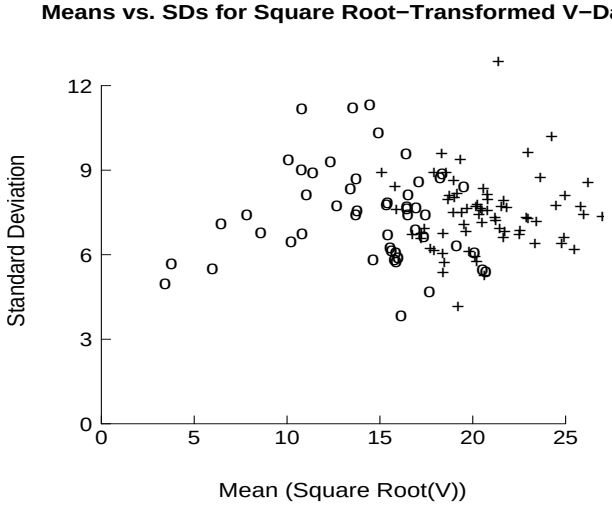
- Given the linear relationship between the moving window standard deviations and means for both the V- and U-data, the natural transformation to try to remove this strong proportional effect ($r_V = 0.78$, $r_U = 0.93$) is a log transformation. More specifically, because there were some 0-values for both variables, the transformation used was a log(V+1)-transformation. This resulted in the moving window means vs. SDs scatterplots shown below. What happened here? Why didn't the log transform remove the proportional effect?

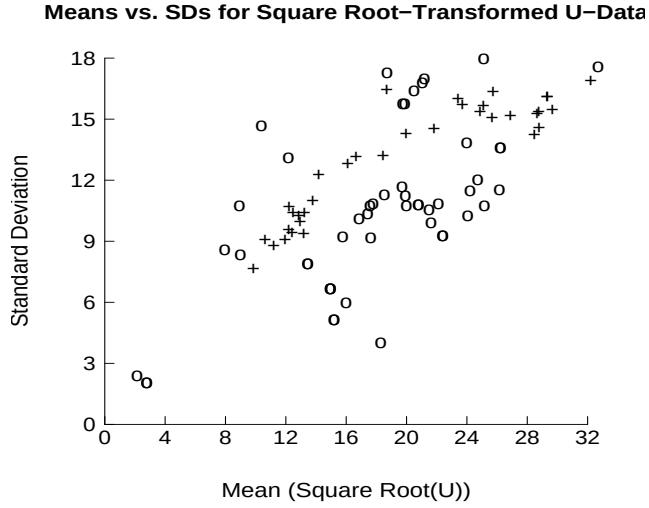




- Often, when the relationship between the means and SDs is linear, there is also a linear relationship between the means and *variances*. If this were true, a square root transform would be the natural choice. Without showing the plots, the correlations between the moving window means and variances for the raw data were: $r_V = 0.73$ and $r_U = 0.92$. These do not differ much from those with the standard deviations.

- This discussion points out the need to consider a number of transformations when trying to remove a proportional effect - it is not as simple as following a recipe. The square root transform was used for V & U, resulting in the two moving window scatterplots shown below. Did this do a better job removing the mean-variance relationship?





Global & Point Estimation (5.4.1, 5.4.2 - Bailey & Gatrell)

We have spent the first few weeks of this course examining various exploratory tools for spatial data. The most common method of modeling geostatistical data involves a description of the covariance structure through the use of a variogram followed by a prediction technique known as kriging. Before following this path, we first discuss some estimation techniques which are based strictly on distances between the observed points in some spatial domain. In general, there are two types of predictions made over some spatial region, as given in the example below.

Example: Suppose you are sampling ore for gold content at n locations $\boldsymbol{s_1, \ldots, s_n}$, where the corresponding response variables are $Y_1, \ldots, Y_n$ with observed outcomes $y_1, \ldots, y_n$. The objective in such a study might be to:

1. Estimate the average gold content in a particular area or over the whole region.
2. Predict the gold content at a new location $\boldsymbol{s_0}$.

Hence, there are two types of predictions made, although the latter type is the more common in practice:

1. Global Estimation: estimating some feature (e.g.: the mean) over the region of interest.
2. Point Estimation: prediction at a specific site.

Although many of the concepts regarding making global (mean) or point estimates are the same, there is one fundamental difference. Point estimation accounts directly for the distances of separation between points in determining weights and making predictions.

Example: Consider the Walker Lake V-concentration data for the sample data. Recall that the 470 measurements were collected in two stages, where the first 195 were taken over a fairly regular grid and the remaining 275 were preferentially clustered at larger initial values.

- Some basic EDA gave a sample mean concentration for these data of $\overline{v} = 436.5$ which turned out to be a *terrible* estimate of the true mean of the exhaustive data set ($\mu = 276.9$). Why?
- What would be a better way of estimating μ?

Weighted Averages: Suppose as in the example above, we observe the values $y_1, \ldots, y_n$ at n sites $\boldsymbol{s_1, \ldots, s_n}$. One general way to make a prediction (either global or at a particular point) is to use a weighted average of the n observed values:

$$\sum_{i=1}^{n} w_i y_i, \quad \text{where:}$$

1. For global estimation, more weight is given to sites separated by larger distances from other sites. In this way, sites which are preferentially clustered are downweighted, as they contain essentially the same information.

2. For point estimation, more weight is given to sites which are closer to the prediction site.

In general, there are many possible ways of weighting the samples, all leading to different estimators. This handout will explore some of these.

One common criterion for an estimator to be a good one is *unbiasedness*. Viewing the data values we observe, $y_1, \ldots, y_n$, as one possible realization from a population of possible outcomes, then we would like whatever estimation method we use to produce predictions whose average coincides with the *expected value* for that site. For any particular realization, the prediction will most likely differ from the expected value, but the average error over all possible realizations should be zero. [It should be noted that this "assumption" of zero average error is fundamental to both multiple linear regression and ANOVA models.]

Some Notation: As a quick review/introduction to some basic ideas in statistical estimation, consider the following:

- The *prediction at site* $\boldsymbol{s_0}$ is: $\widehat{Y}(\boldsymbol{s_0}) = \sum_{i=1}^{n} w_i Y(\boldsymbol{s_i})$ where the w_i's are weights.
- The *estimation (or prediction) error at site* $\boldsymbol{s_0}$ is: $\widehat{\epsilon}(\boldsymbol{s_0}) = Y(\boldsymbol{s_0}) - \widehat{Y}(\boldsymbol{s_0})$.
- The *expectation of the estimation error*, $E\left[\widehat{\epsilon}(\boldsymbol{s}_0)\right]$, is found as follows. First, we compute the expectation of the prediction at site $\boldsymbol{s_0}$:

$$E(\widehat{Y}(\boldsymbol{s_0})) = E\left[\sum_{i=1}^{n} w_i Y(\boldsymbol{s_i})\right] = \sum_{i=1}^{n} w_i E(Y(\boldsymbol{s_i})).$$

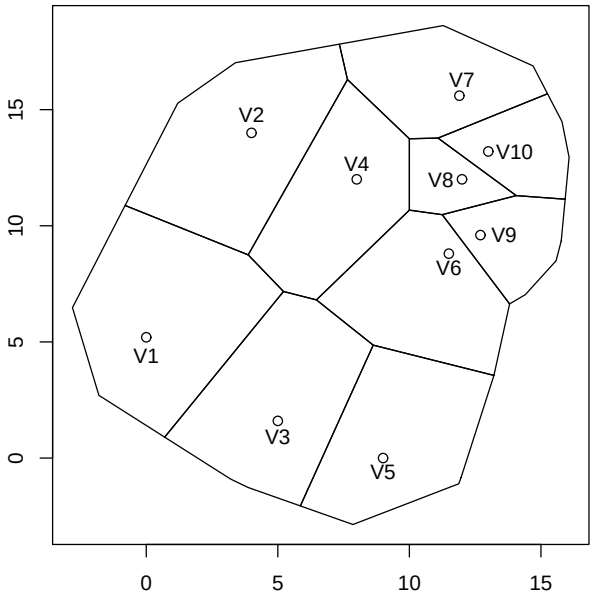
Under either intrinsic or 2nd-order stationarity, we know that: $E(Y(\boldsymbol{s_i})) = \mu$ for all i. What are the consequences of this for the estimator $\widehat{Y}(\boldsymbol{s_0})$?

The condition $\sum w_i = 1$ is known as the "unbiasedness condition." Weighted averages of the observed values produce unbiased predictions if and only if the sum of the weights is one.

Global Estimation - Estimating the mean over the region

The objective here is to estimate the mean response over some region of interest. For this type of estimation, we can use a type of weighted average. Here, we examine two different weighting schemes for estimating the global mean.

1. **Polygonal Declustering**: This method uses *polygons of influence* (also called Voronoi polygons) to determine the area of influence of a sampled point. To see how this works, consider the sampled points shown to the right:



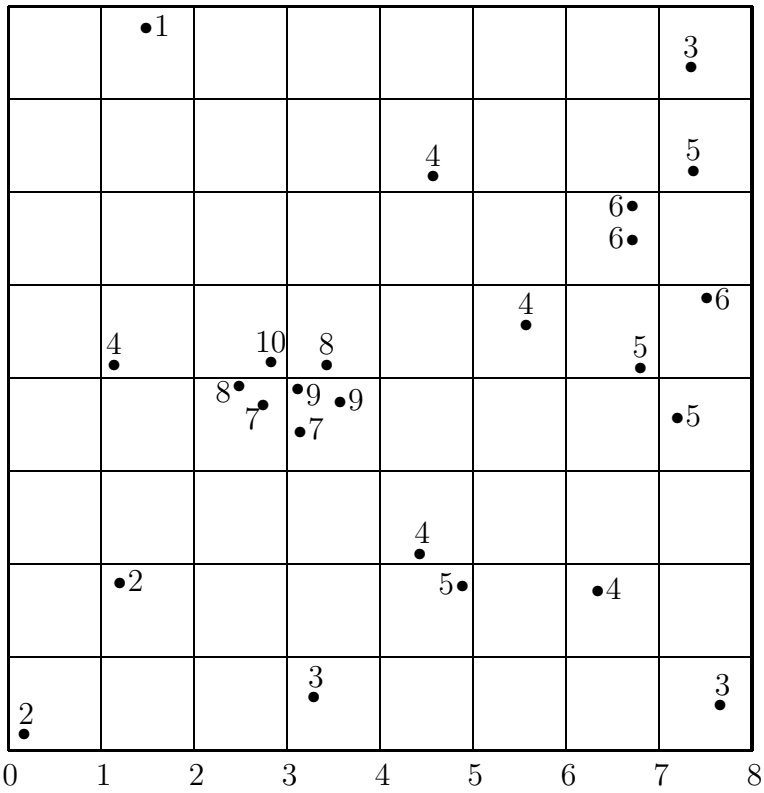
The global mean is then estimated by taking a weighted average over the entire region, where the sample values are weighted by the relative areas of their polygons of influence. Notationally, letting y_i = the response at location $\boldsymbol{s_i}$, A_i = the area of the influence polygon for y_i, and $A = \sum A_i$ equal the total area of the influence polygons, the mean for the entire region is estimated by:

$$\overline{y}_{PD} = \sum_{i=1}^{n} \frac{A_i}{A} y_i.$$

Notes:

- Sample locations which are separated by greater distances from other sample locations have larger polygons of influence. This makes sense as such locations represent larger areas of the region and so are entitled to larger weights.
- This type of weighting avoids the problems with unweighted estimation caused by preferential (or clustered) sampling, as seen with the Walker Lake data. Points which are preferentially sampled will be close together (hence providing relatively the same information) and receive less weight when computing the global mean.
- This concept will be revisited under the objective of point estimation, where it will be seen that one method of prediction at any point inside a sample value's polygon of influence uses that sample value as the estimate.
- I wrote a function in **R** called **polydec** which takes as arguments the vectors of s_1- and s_2-coordinates, as well as a "peel" factor (default=1) indicating the degree of tolerance in generating a convex hull around the points, and returns the total area of the convex polygon and the individual polygon areas. Additionally, it generates a graph of the Voronoi polygons like the one shown above. The higher the peel factor (must be an integer), the less likely you are to have edge effects.

2. **Cell Declustering**: This method divides the region into rectangular cells, and samples receive a weight inversely proportional to the # of samples falling in the same cell. To see how this works, consider the spatial layout shown below. The observed values at each site are given next to each point.



Square Cell Size	Estimated Mean
0.1	
1	
2	
4	
8	

To find the cell declustered mean, let:

n = the sample size ($n = 25$ here),
n_i = the # of samples falling in the cell with sample i (including sample i),
d = the number of cells containing at least one sample.

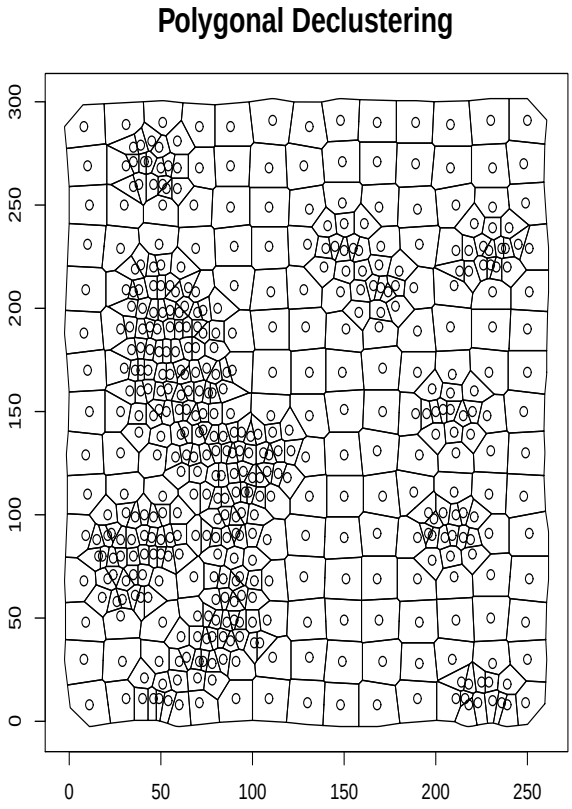
Then the cell declustered mean is given by:

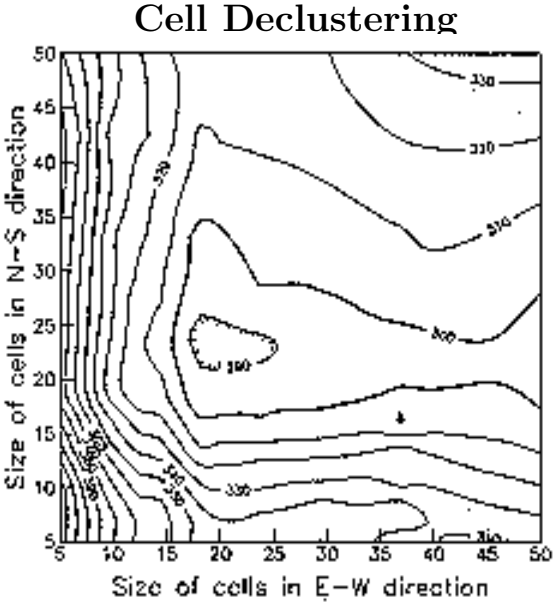
$$\sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} \left[\frac{1}{d n_i}\right] y_i = \frac{1}{d} \sum_{i=1}^{n} \frac{y_i}{n_i}.$$

- The unweighted mean for these data is: $m = \frac{1}{25}\sum y_i = \frac{1}{25}(130) = \underline{5.2}$. In viewing the region above, would you expect this unweighted mean to be biased for the global mean, and if so, in what direction?

- Find the estimated global mean for the various square cell sizes for these data.

Comparison for Walker Lake Data: The polygon of influence map for the Walker Lake V-data is given below. Also given is a contour plot of the cell declustered means for a variety of rectangle sizes. The estimated mean using polygonal declustering was 276.8 and the smallest mean obtained from the cell declustering (using a 20x23 rectangle) was 288 ppm. As the true mean was 276.9, the declustering methods both successfully removed the majority of the effect of the preferential clustering in this case (the raw mean was 436.5!).





- In viewing the cell declustering contours, the minimum sample mean occurs for a rectangular window of size 20x23. Are we justified in choosing the smallest mean here?

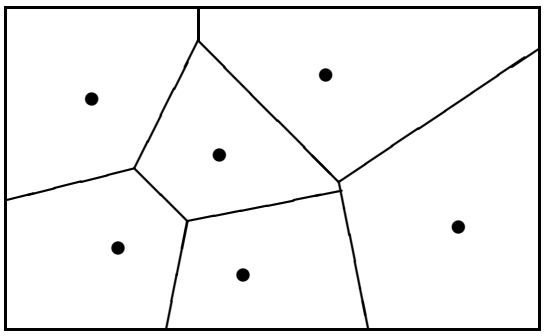
- Note: Points preferentially clustered are severely downweighted.
- One limitation of these declustering methods is that if some areas are scarcely sampled (as with the U-data), prediction will be difficult and likely unreliable.

Point Estimation - Prediction at a Point Not Sampled

The objective here is to predict the value at a location that is not in the sample. For this type of prediction, a weighted average of *neighboring* points will be used. Some of the corresponding weighting schemes (which differ essentially in how they define a neighborhood) are given here.

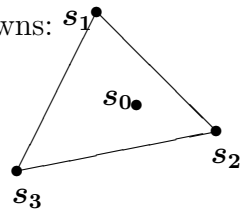
1. **Polygonal Declustering**: This method is an adaptation of the polygonal declustering method used with global estimation, where here a prediction at any new location falling inside a sample point's polygon of influence is given the value of the sample point. In other words, all locations inside a given polygon of influence would receive the same estimate. Once you move to a site in a different polygon, the estimate jumps to whatever the sample value is that generated the new polygon. Summarizing then: the polygonal declustering method chooses the closest point as the estimate.

   This type of estimation can be viewed as a weighted average where the closest sampled point is given a weight of $w_i = 1$, and all other points are given a weight of $w_i = 0$.



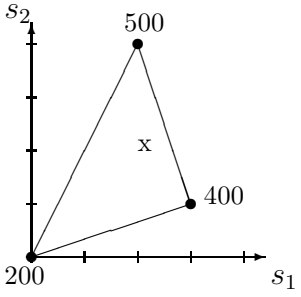
2. **Triangulation**: The method of triangulation fits a plane through three nearby samples. The predicted value is the value of the plane at the coordinates (s_{10}, s_{20}) of the prediction site. To see how to fit this plane, suppose we have 3 points, say (s_{11}, s_{21}), (s_{12}, s_{22}), (s_{13}, s_{23}), that surround the location for which a prediction is wanted. Fitting the plane amounts to solving the system of three equations in three unknowns:

$$\begin{bmatrix} s_{11} & s_{12} & 1 \\ s_{21} & s_{22} & 1 \\ s_{31} & s_{32} & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \Rightarrow \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & 1 \\ s_{21} & s_{22} & 1 \\ s_{31} & s_{32} & 1 \end{bmatrix}^{-1} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}.$$



   The resulting predicted value at location $\boldsymbol{s_0} = (s_{01}, s_{02})$ is: $\widehat{y}_0 = as_{01} + bs_{02} + c$.

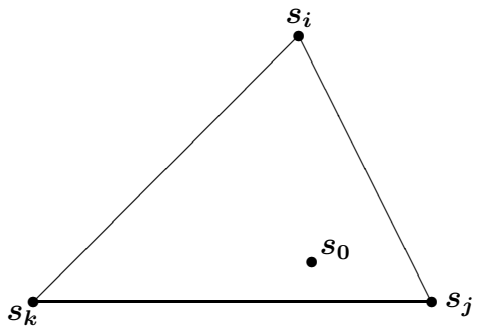
   Example: Suppose we have values 200, 400, & 500 located at locations (0,0), (3,1), & (2,4) respectively, and want to predict the value at location (2,2).



- This method avoids the discontinuities inherent with the polygonal declustering method. Such discontinuities are generally undesirable as they are generally not a realistic portrayal of the continuous nature of most spatial data.
- The triangulation estimate can also be viewed as a weighted average of the three sample site values y_i, y_j, y_k. The weights in this case are the areas of the opposite triangles. That is, if we form a triangle of the three sample points surrounding $\boldsymbol{s_0}$, then three subtriangles are formed with the point $\boldsymbol{s_0}$. The resulting weighted average is given by:

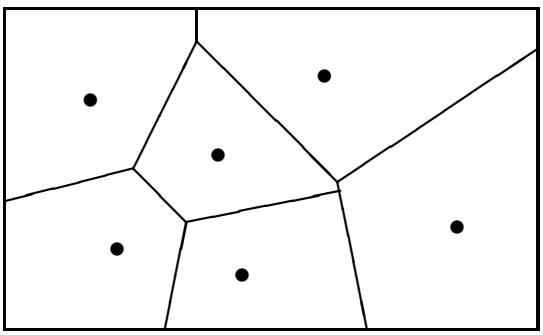
$$\widehat{y}_0 = \frac{A_{0jk}y_i + A_{0ik}y_j + A_{0ij}y_k}{A_{ijk}}, \text{ where:}$$

A_{rst} = the area of the triangle formed by points r, s, t.



When forming triangles, it is best to use triangles that do not have any very small or very large angles, because long thin regions could traverse very different types of terrain. The idea is to make compact triangles where the values will be fairly similar to one another.

- **Delaunay Triangulation**: This type of triangulation, based on the area of influence polygons, generates triangles that have the smallest possible distances and are as equilateral as possible. To see how this works, consider the points below.
  - First, construct the area of influence polygons for each sample point in this region. These polygons are often referred to as Voronoi polygons.
  - If 3 points share a common vertex of influence polygons, they form a Delaunay triangle. The resulting triangles are as close to equilateral as possible.
  - The **R** function **delaunay** (library **spatstat**), or **tri.mesh** (library **tripack**), will plot the Delaunay triangles for a given set of locations. For (x, y)-coordinates, you can use the **delaunay** function by typing:



```
xy.ppp <- ppp(x,y,range(x),
              range(y))
out <- delaunay(xy.ppp)
plot(out)
```

Some Final Comments on Triangulation

- Triangulation should not be used for extrapolation purposes (i.e.: each triangle should only be used to make predictions at locations within that triangle).
- Because we are using closed regions to make predictions, it is often not possible to use triangulation to predict around the edges of a region. Again, edge effects with all types of spatial data are often a problem.

3. Inverse Distance Methods: These methods of prediction give more weight to the closest samples and less weight to those that are far away. Hence, the weights themselves depend on the *distance* from samples to the prediction site. Within the class of inverse distance methods, there are several types.

   (a) Standard Method: Use a weighted average of *all* sample points with weight inversely proportional to distance. This gives the following estimator:

   $$\widehat{y}_0 = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \quad \text{where } w_i = \frac{1}{d_i} \text{ and } d_i \text{ is the distance from } \boldsymbol{s_i} \text{ to } \boldsymbol{s_0}.$$

   (b) Local Sample Mean: Specify in some way all samples which are "nearby" the prediction site, and take a weighted average of these m sites via:

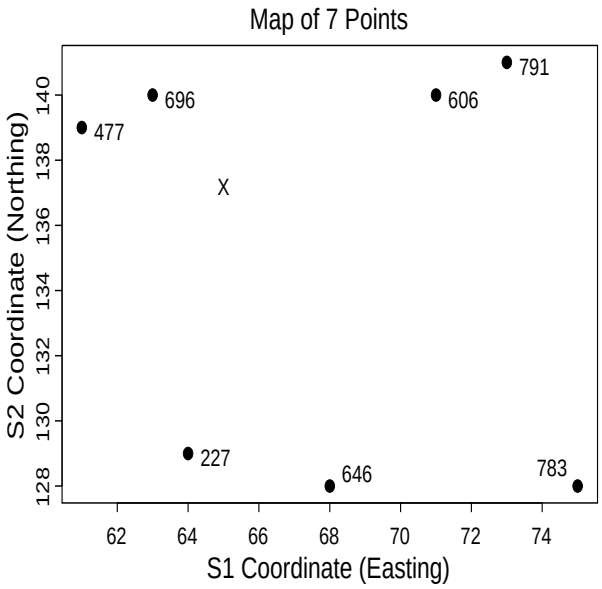
   $$\widehat{y}_0 = \frac{\sum_{i=1}^{m} w_i y_i}{\sum_{i=1}^{m} w_i}, \quad \text{where } w_i = \frac{1}{d_i} \text{ and } d_i \text{ is the distance from } \boldsymbol{s_i} \text{ to } \boldsymbol{s_0}.$$

   This alternative to the standard inverse distance method allows for our belief that beyond some range of influence, points have essentially no effect on the prediction site. The most common choice is an unweighted local sample mean, which is just unweighted m-nearest neighbor estimation.

   (c) Power Alternative: Weights can also be chosen so that they are inversely proportional to some power of the distance, i.e.: $w_i = 1/d_i^p$.

   - When $p = 0$, the weights are all 1, so every point gets equal weighting and we estimate $y(\boldsymbol{s_0})$ by the arithmetic mean.
   - As p increases, we give more and more weight to nearby samples and the contribution of those far away approaches zero. A good example of this is shown at the top of the next page.
   - The most common choice for p is $p = 2$ (inverse distance squared weighting).
   - Different choices of power p can be applied to either of the two methods given above (Standard Method or Local Sample Mean).

Example: To illustrate the use of these weighting techniques, consider the local neighborhood around the point (65,137) in the Walker Lake sample data, as shown to the right. We might assume that these 7 points are all points within some distance of the prediction site (65,137) or that we have chosen to look at the 7 nearest neighbors to make a prediction.



The different estimation methods used are summarized in the following tables.

Inverse Distance Weighting Calculations for Sample Values in Vicinity of 65E, 137N

	Sample Number	s_1	s_2	V	Distance from 65E, 137N	$1/d_i$	$\frac{1/d_i}{\sum 1/d_i}$
1	225	61	139	477	4.5	0.2222	0.2088
2	437	63	140	696	3.6	0.2778	0.2610
3	367	64	129	227	8.1	0.1235	0.1160
4	52	68	128	646	9.5	0.1053	0.0989
5	259	71	140	606	6.7	0.1493	0.1402
6	436	73	141	791	8.9	0.1124	0.1056
7	366	75	128	783	13.5	0.0741	0.0696
					$\sum 1/d_i =$	1.0644	

Effect of Inverse Distance Exponent on Sample Weights and on V-Estimate

		$\frac{1/d_i^p}{\sum 1/d_i^p}$					
	V	$p = 0.2$	$p = 0.5$	$p = 1.0$	$p = 2.0$	$p = 5.0$	$p = 10.0$
1	477	0.1564	0.1700	0.2088	0.2555	0.2324	0.0106
2	696	0.1635	0.1858	0.2610	0.3993	0.7093	0.9874
3	227	0.1390	0.1343	0.1160	0.0789	0.0123	<.0001
4	646	0.1347	0.1260	0.0989	0.0573	0.0055	<.0001
5	606	0.1444	0.1449	0.1402	0.1153	0.0318	0.0019
6	791	0.1364	0.1294	0.1056	0.0653	0.0077	<.0001
7	783	0.1255	0.1095	0.0696	0.0284	0.0010	<.0001
$\hat{v}$ (in ppm)		601	598	594	598	637	693

- As $p \to \infty$, the estimate based on inverse distance weighting approaches that of polygonal declustering.

With all of these estimation methods in place, the next step will be to identify means of evaluation for these estimators. Before doing that, this seems to be the proper place to introduce the notion of kriging in a non-mathematical context. Any decent estimation procedure for spatial data needs to account for the spatial correlation (continuity) present in some fashion (otherwise, we are assuming independence of the site values). The estimation methods in this handout account for spatial continuity in a variety of ways.

- Polygonal declustering uses the closest site as the estimate.
- Triangulation uses three nearby sites with a planar interpolation.
- Inverse distance weighting improves upon these ideas by taking into account the Euclidean distance between samples and the prediction site.
- In all of these methods, there is one vital aspect of the data which is being neglected. What is it?

- It is frequently debated whether the added complexity involved in modeling the covariance structure with variograms is worth the effort. In fact, some might argue that since the specification of a variogram model is certainly subject to error, then using such a model may actually *add* bias to predictions and hence be worse than simpler distance-based methods. What do you think? Can you think of cases where accounting for spatial correlation is likely to be unnecessary?