University of California, Los Angeles
Department of Statistics

Statistics C173/C273 **Instructor: Nicolas Christou**

Spatial statistics

- Why spatial statistics? Noel Cressie ("Statistics for Spatial Data") writes "why, how, when" are not enough. We need to add "where".

- Today, spatial statistics models appear in areas such as mining, geology, hydrology, ecology, environmental science, medicine, image processing, crop science, epidemiology, forestry, atmospheric science, etc.

- Need to develop models that deal with data collected from different spatial locations.

- The basic components are the spatial locations $\{s_1, s_2, \cdots, s_n\}$ and the data observed at these locations denoted as $\{Z(s_1), Z(s_2), \cdots, Z(s_n)\}$.

- The distance between the observations is important in analyzing spatial data. With distance we mostly mean "Euclidean distance". However there are other forms of distances (e.g. road miles, travel time, etc.). The latter is modeled through multidimensional scaling. Here we will consider mostly (if not always) Euclidean distances.

- Consider the following example taught in all introductory statistics courses:
  Let the spatial data $Z(s_1), Z(s_2), \cdots, Z(s_n)$ be an i.i.d. sample from $N(\mu, \sigma_0)$. The MVUE of μ is

  $$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z(s_i)$$

  We know that $\bar{Z} \sim N(\mu, \frac{\sigma_0}{\sqrt{n}})$, and therefore we can easily construct a 95% confidence interval for μ as follows:

  $$\bar{Z} \pm 1.96\frac{\sigma_0}{\sqrt{n}}$$

- The previous example assumes an i.i.d. sample. This can be too simplistic for spatial data. A more realistic assumption is that the data exhibit some spatial correlation. Suppose this spatial correlation is represented through the covariance function

  $$cov(Z(s_i), Z(s_j)) = \sigma_0^2 \rho^{|i-j|}$$

  In the i.i.d. case $cov(Z(s_i), Z(s_j)) = 0$ (independent therefore the covariance is zero).

- What is the confidence interval for the non-i.i.d. case? Find the variance of $\bar{Z}$ first:

$$var(\bar{Z}) = var\left(\frac{1}{n}\sum_{i=1}^{n} Z(s_i)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(Z(s_i), Z(s_j)\right)$$

  Or after some simplification ...

$$var(\bar{Z}) = \frac{\sigma_0^2}{n}\left[1 + 2\left(\frac{\rho}{1-\rho}\right)\left(1 - \frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2\left(\frac{1-\rho^{n-1}}{n}\right)\right]$$

- Since $Z(s_i)$ is Gaussian the distribution of $\bar{Z}$ is also normal with mean μ and standard deviation the square root of the above expression.

- Suppose $n = 10$ and $\rho = 0.26$. For this example we get:

$$var(\bar{Z}) = \frac{\sigma_0^2}{10}(1.608) \;\Rightarrow \bar{Z} \sim N\left(\mu, \frac{\sigma_0\sqrt{1.608}}{\sqrt{10}}\right)$$

  and a two-sided 95% confidence interval for μ is

$$\bar{Z} \pm 1.96\frac{\sigma_0\sqrt{1.608}}{\sqrt{10}} \quad \text{or} \quad \bar{Z} \pm 2.485\frac{\sigma_0}{\sqrt{10}}$$

- Conclusion: If we do not realize the presence of spatial correlation in our data and we use $\bar{Z} \pm 1.96\frac{\sigma_0}{\sqrt{10}}$, we obtain a confidence interval that is *too narrow*. The actual coverage is 87.8%, not 95%. Why?

- Linear models with spatially dependent error term:
  The classical regression model in matrix form when the error terms are i.i.d. random variables is given by

$$Z = X\beta + \epsilon, \text{ with } var(\epsilon) = \sigma^2 I$$

  and the estimation of β is obtained through OLS

$$\hat{\beta}_{ols} = (X'X)^{-1}X'Z$$

  When the error terms are spatially correlated the above model can be written as

$$Z = X\beta + \delta, \text{ with } var(\delta) = \Sigma$$

  where Σ is the $n \times n$ variance-covariance matrix. The form of Σ usually is not known. If Σ is known the estimation of β could be obtained through generalized least squares (gls)

$$\hat{\beta}_{gls} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Z$$

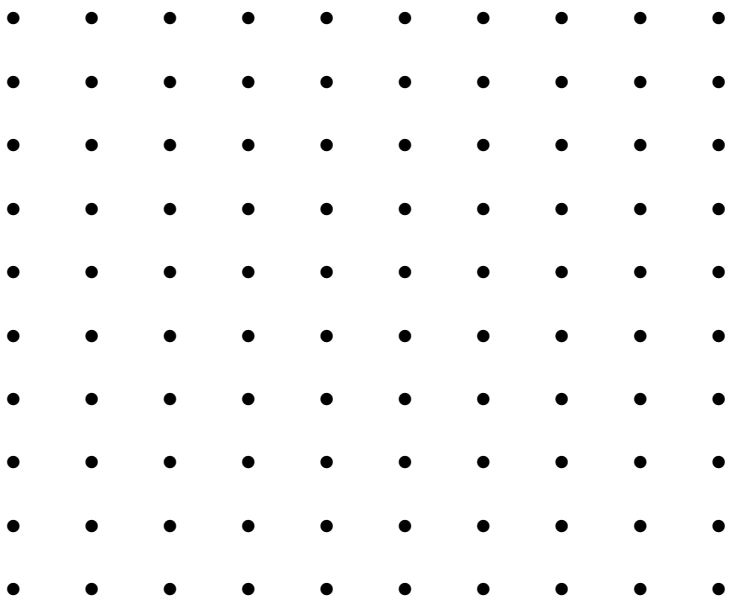
Geostatistics

- Matheron (1963) develop a comprehensive theory mainly for spatial prediction for the estimation of ore reserves.
- Geostatistics can be applied to data sets not only pertaining the earth sciences.
- The data are thought as a realization of

$$Z(s) : s \in D$$

- The data are denoted with $\{Z(s_1), Z(s_2), \cdots, Z(s_n)\}$ and are collected at locations $\{s_1, s_2, \cdots, s_n\}$.
- The cornerstone of geostatistics is the *variogram* which describes the spatial correlation.
- First let's discuss the so called "h-scatterplots".

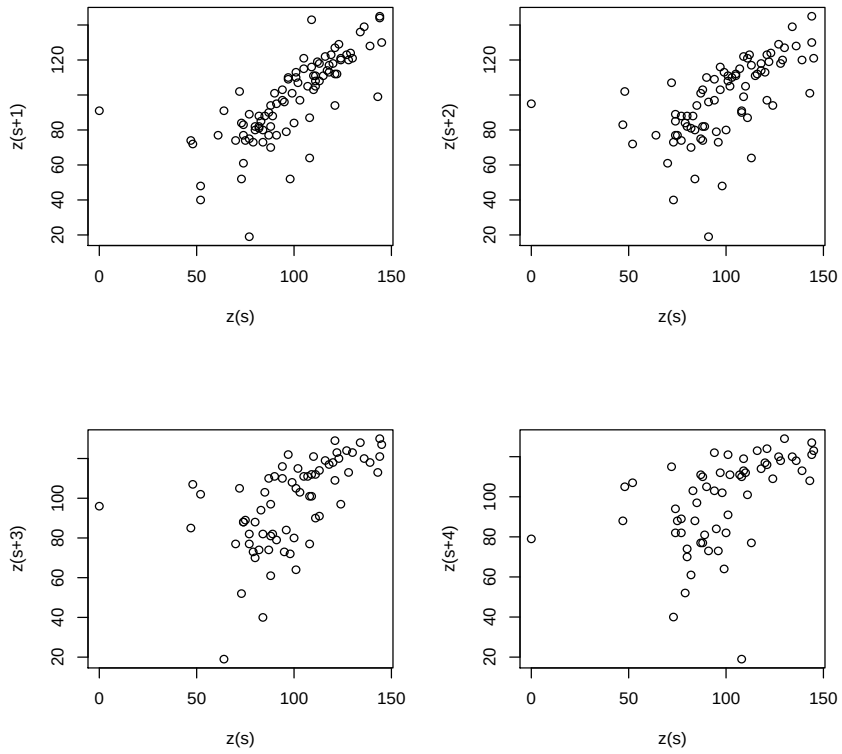
  An h-scatterplot shows all possible pairs of data values whose locations are separated by a certain distance in a particular direction. For example, the figure below shows a 10×10 grid (100 data points - with distance between two consecutive data points 1 m in north-south and east-west direction).



  And on the next page we see the value at each location.

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	96	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

In this first example we will plot each value (we place it on the x-axis) against a value that is 1 m apart on the south-north direction (we place it on the y-axis). How many pairs are there? There are $9 \times 10 = 90$ pairs. The scatterplot shows a cloud of points distributed around the 45-degree line. We observe some similarity between nearby data points. If we increase the distance from 1 m to 2 m, 3 m, and 4 m we see that the cloud becomes "fatter" indicating that the values separated by longer distance are not as close as with the 1 m case. These plots are shown below:



- Similarly, if we move on southwest-northeast direction we will pair all values that are $\sqrt{2}\ m$ apart. In this case there are 81 pairs.

- As with many graphical displays, we need a quantitative summary of the information contained on an h-scatterplot. In the previous figure we can measure the "fatness" of the points by the correlation coefficient. The correlation coefficient decreases as the separation distance increases. The relation between the correlation coefficient and h (the separation distance) it is called *correlogram*. We can also plot the covariance against h to get the so called *covariogram*.

- How do we compute the covariance and the correlation coefficient for data points separated by h?

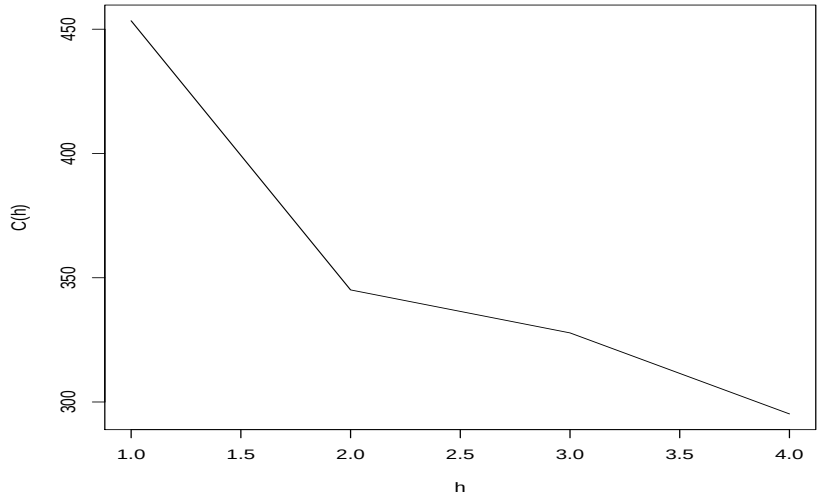
$$C(h) = \frac{1}{N(h)} \sum_{h_{ij}=h} (Z_i - \bar{Z}_s)(Z_j - \bar{Z}_{s+h})$$

$$\rho(h) = \frac{C(h)}{\sigma_{z_s} \sigma_{z_{s+h}}}$$

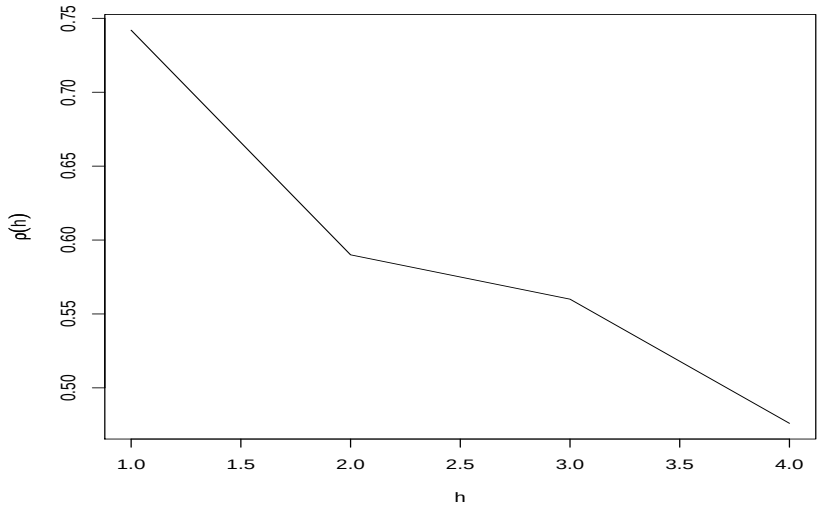
where $\sigma_{z_s}, \sigma_{z_{s+h}}$ are the standard deviations of the variable Z_s and the variable Z_{s+h}.

- The two figures below show the covariogram and correlogram for these data.

Covariogram:



Correlogram:



h	$C(h)$	$\rho(h)$
1	453.4	0.742
2	345.1	0.590
3	327.8	0.560
4	295.2	0.476

- You can verify the result for $h = 1$ using the 90 paired values below:

Z(s)	Z(s+1)	Z(s)	Z(s+1)
87	77	109	143
77	75	143	99
75	74	99	101
74	77	101	113
77	89	113	108
89	88	108	64
88	82	64	91
82	82	91	77
82	81	77	19
100	84	0	91
84	80	91	95
80	80	95	96
80	82	96	79
82	88	79	73
88	70	73	84
70	74	84	73
74	61	73	52
61	77	52	40
47	74	98	52
74	83	52	48
83	85	48	72
85	88	72	102
88	94	102	107
94	103	107	105
103	97	105	115
97	110	115	111
110	103	111	111
111	108	134	136
108	87	136	139
87	90	139	128
90	101	128	120
101	110	120	118
110	111	118	113
111	105	113	118
105	121	118	117
121	112	117	114
124	121	144	144
121	94	144	145
94	97	145	130
97	109	130	121
109	116	121	127
116	122	127	123
122	112	123	129
112	119	129	124
119	123	124	120

Constructing h-scatterplots using the library `gstat`:
Let's use only the first three rows of the data set above. Again the distance on the north-south and east-west direction is 1 m.

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

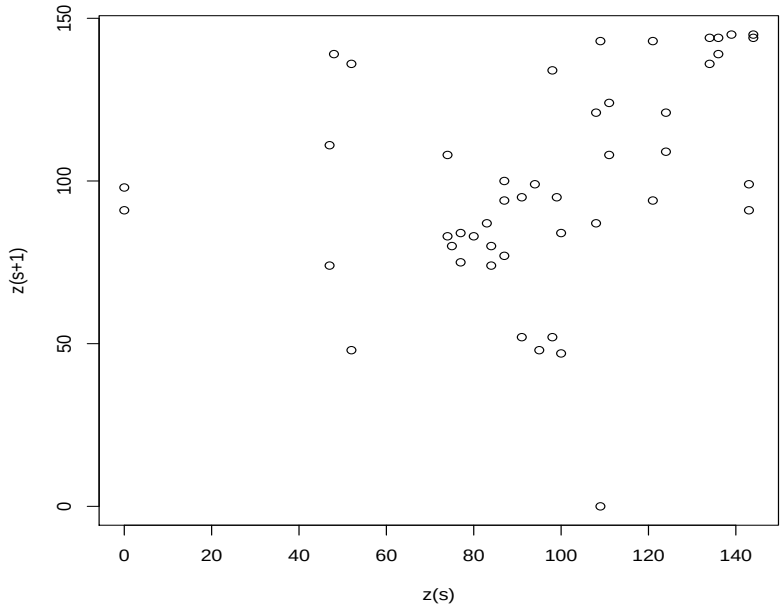
Here are the data with their coordinates:

```
   x y   z
1  0 0  87
2  1 0 100
3  2 0  47
4  3 0 111
5  4 0 124
6  5 0 109
7  6 0   0
8  7 0  98
9  8 0 134
10 9 0 144
11 0 1  77
12 1 1  84
13 2 1  74
14 3 1 108
15 4 1 121
16 5 1 143
17 6 1  91
18 7 1  52
19 8 1 136
20 9 1 144
21 0 2  75
22 1 2  80
23 2 2  83
24 3 2  87
25 4 2  94
26 5 2  99
27 6 2  95
28 7 2  48
29 8 2 139
30 9 2 145
```

Here are all the pairs separated by distance $h = 1$ north-south and east-west:

	z(s)	z(s+1)
1	87	77
2	77	75
3	100	84
4	84	80
5	47	74
6	74	83
7	111	108
8	108	87
9	124	121
10	121	94
11	109	143
12	143	99
13	0	91
14	91	95
15	98	52
16	52	48
17	134	136
18	136	139
19	144	144
20	144	145
21	87	100
22	100	47
23	47	111
24	111	124
25	124	109
26	109	0
27	0	98
28	98	134
29	134	144
30	77	84
31	84	74
32	74	108
33	108	121
34	121	143
35	143	91
36	91	52
37	52	136
38	136	144
39	75	80
40	80	83
41	83	87
42	87	94
43	94	99
44	99	95
45	95	48
46	48	139
47	139	145

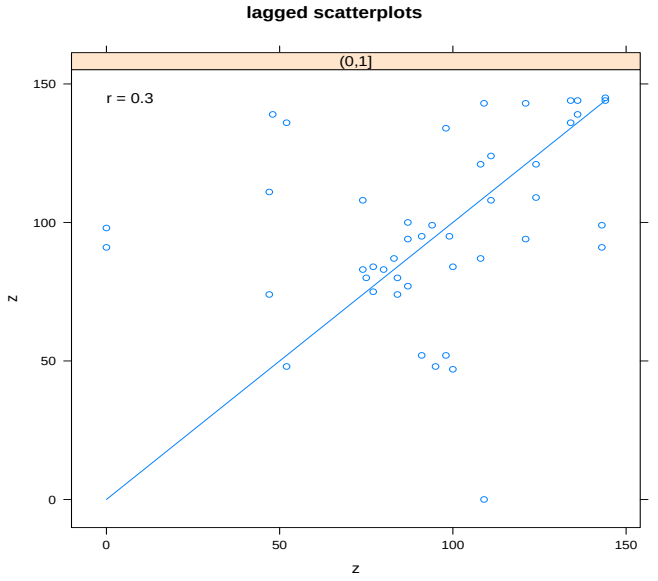
The $h-$scatterplot of $z(s)$ against $z(s+1)$ is shown below:

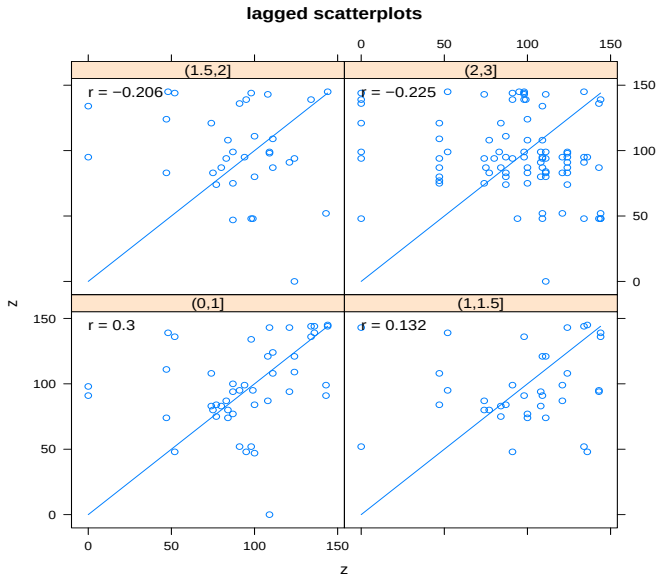


Using the library `gstat` you can produce $h-$scatterplots as follows:

```
a <- read.table("h_scatter_1.txt", header=T)  #Read the data (page 8 above)
coordinates(a) <- ~x+y  #Convert the data into spatial data
hscat(z~1, a, c(0,1))   #This will produce the z(s) against z(s+1) plot.
hscat(z~1, a, c(0,1,1.5, 2,3))   #This will produce 4 different
#h-scatterplots with h=1, 2^0.5, 2, 3.
```

Here they are:





The variogram

Suppose that the vector Z of the observed values at spatial locations is a realization of a random process $Z(s) : s \in D$. Then intrinsic stationarity is defined as follows:

$$E(Z(s+h) - Z(s)) = 0$$

and

$$Var(Z(s+h) - Z(s)) = 2\gamma(h)$$

The quantity $2\gamma(h)$ is known as the variogram and is very crucial in geostatistics. The variogram says that differences of variables lagged h-apart vary in a way that depends only on h through the length of h. This is called *isotropic* variogram as opposed to *anisotropic* variogram which depends not only on the length h but also the direction. Because of the assumption of constant mean (no trend) we have $E(Z(s)) = \mu$ and we can write

$$\text{Var}(Z(s+h) - Z(s)) = E(Z(s+h) - Z(s))^2 = 2\gamma(h)$$

Therefore we can use the method of moments estimator for the variogram (also called the classical estimator):

$$2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{N(h)} (Z(s_i) - Z(s_j))^2,$$

where the sum is over $N(h)$ such that $s_i - s_j = h$.

- The following is the so called semivariogram.

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{N(h)} (Z(s_i) - Z(s_j))^2,$$

- Important: The values of $C(h), \rho(h), \gamma(h)$ are unaffected if we switch all he i, j subscripts. For example, south-north or north-south. Therefore:

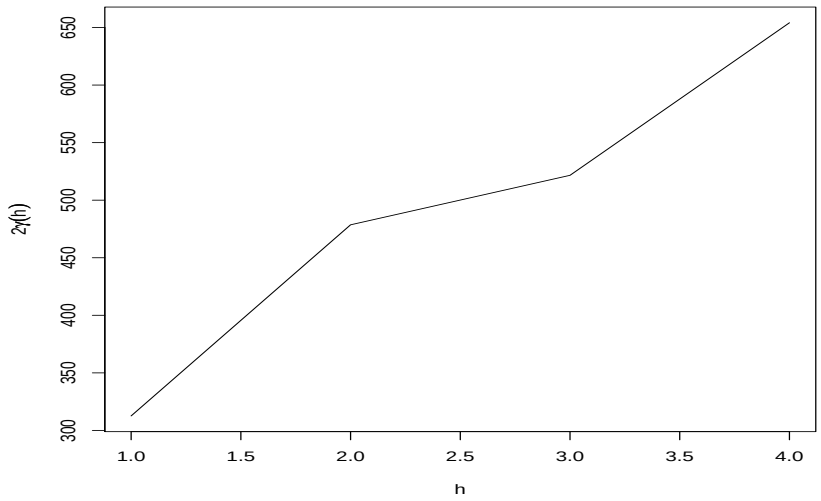
$$C(h) = C(-h),\ \rho(h) = \rho(-h),\ \gamma(h) = \gamma(-h)$$

- The table below gives the values of the semivariogram for the values of $h = 1, 2, 3, 4$ meters on the south-north direction.

h	Variogram
1	312.6
2	478.6
3	521.6
4	654.1

- For our data the variogram is shown below. We observe that it increases as h increases.

  Variogram:



- Notes:

  - The variogram can be calculated for any direction.
  - Our goal is to estimate every single point on a dense grid of the area of interest (and produce a raster map). Therefore we want to calculate the variogram for many values of the lagged distance (h). In our example, what if we want to pair data points which are 1.3 m apart? Based on the geometry of our data there isn't such a distance. This is solved if we can specify tolerances for both h and direction. We can use a tolerance of ± 0.5 m and ± 20 degrees. Using these tolerances we will be able to produce enough pairs.

  – The semivariogram now is calculated as follows:

  $$\hat{\gamma}(h) = \frac{1}{2N(\approx h)} \sum_{N(h)} (Z(s_i) - Z(s_j))^2,$$

  In this formula we choose to sum over the pairs that are approximately h apart.

  – Same applies for irregular sampling. To compute the variogram when $h = 100$ m with a tolerance of 20 m we will pair all the observations that fall between $80 - 120$ m.

  – It is a good practice to calculate directional variograms (unless we are sure that there is no anisotropy in our data).

- Another example:
  Using the data below, calculate and plot the semivariogram for the following directions and separation distances (h):

  - North-south ($h = 100, 200, 300, 400, 500$).
  - East-west ($h = 100, 200, 300, 400, 500$).
  - Northeast-southwest ($h = 141.42, 282.84, 424.26$).
  - Northwest-southeast ($h = 141.42, 282.84, 424.26$).

  The vertical and horizontal distance between two points is 100 m.

44		40	42	40	39	37	36	
42		43	42	39	39	41	40	38
37	37	37	35	38	37	37	33	34
35	38		35	37	36	36	35	
36	35	36	35	34	33	32	29	28
38	37	35		30		29	30	32

Assumptions: Stationarity

- **Stationarity of order 2:**
  A random function is said to be stationary of order 2 if

  1. The expected value of $Z(s)$ does not depend on point s. Therefore,

     $$E(Z(s)) = \mu, \ \forall \ s$$

  2. For each pair of random variables $Z(s)$ and $Z(s+h)$ the covariance exists and depends only on the separation distance h:

     $$cov[Z(s+h), Z(s)] = C(h) = E(Z(s) - \mu)(Z(s+h) - \mu) \Rightarrow$$

     $$C(h) = E\left[Z(s)Z(s+h)\right] - \mu^2$$

  3. As a result of the above the following is true:

     $$Var(Z(s)) = E\left[Z(s) - \mu\right]^2 = C(0), \ \forall \ s$$

     $$\gamma(h) = \frac{1}{2}E\left[Z(s+h) - Z(s)\right]^2 = C(0) - C(h), \ \forall \ s$$

     It follows that under the second-order stationarity assumption the covariance and the variogram are two equivalent tools for describing the correlation between the random variables $Z(s+h)$ and $Z(s)$ separated by a distance h.

  4. Another tool is the correlogram that follows directly from the definition $\rho = \frac{Cov}{\sigma_1 \sigma_2}$:

     $$\rho(h) = \frac{C(h)}{C(0)} = \frac{C(0) - \gamma(h)}{C(0)} = 1 - \frac{\gamma(h)}{C(0)}$$

- **Intrinsic stationarity:**
  The second-order stationarity assumes that the covariance exists. But there are cases where the covariance and the variance do not exist. However in these cases the variogram does exist! Therefore, intrinsic stationarity is defined as follows:

  1. The expectation does not depend on s:

     $$E(Z(s)) = \mu, \ \forall \ s$$

  2. For all separation distances h the variance of the difference $Z(s+h) - Z(s)$ does not depend on s:

     $$Var\left[Z(s+h) - Z(s)\right] = E\left[Z(s+h) - Z(s)\right]^2 = 2\gamma(h)$$

     We observe that second-order stationarity (a stricter form of stationartity) implies intrinsic stationarity but not the other way around!

Properties of the covariance and the variogram

1. **Covariance:**
   Let $Z(s_1), Z(s_2), \cdots, Z(s_n)$ be stationary random variables with expectation μ and covariance $C(h)$. Let Y be a linear combination of these variables, i.e.

   $$Y = \sum_{i=1}^{n} \omega_i Z(s_i)$$

   Aside note: we will consider such linear combinations later in kriging.

   This linear combination is a random variable itself and must have nonnegative variance

   $$Var(Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j C(s_i - s_j) \geq 0$$

   Therefore the covariance function must ensure that the variance of Y is always nonnegative. Such a function it is called "positive definite" and this is a property that must be possessed by the covariance function.

2. **Variogram:**
   Using the expression

   $$\gamma(h) = \frac{1}{2} E\left[Z(s+h) - Z(s)\right]^2 = C(0) - C(h), \ \forall \ s$$

   the variance of Y in terms of the variogram can be expressed as follows:

   $$Var(Y) = C(0) \sum_{i=1}^{n} \omega_i \sum_{j=1}^{n} \omega_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j \gamma(s_i - s_j)$$

   In the case where the variance $C(0)$ does not exist the expression above becomes

   $$Var(Y) = -\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j \gamma(s_i - s_j)$$

   This is true on the condition that $\sum_i^n \omega_i = 0$. Again, because the variance of Y must be nonnegative, the variogram must satisfy the property of conditional negative definiteness:

   $$\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j \gamma(s_i - s_j) \leq 0$$

Some other properties:
In practice, the covariance between two random variables $Z(s+h)$ and $Z(s)$ disappears as h becomes too large.

$$C(h) \to 0, \quad \text{when} \quad h \to \infty$$

We set $C(h) = 0$ when $h \geq \alpha$ where α is the so called range (see figure below).

For the semi-variogram using

$$\gamma(h) = C(0) - C(h), \ \forall \ s$$

we observe that:

$$\gamma(\infty) = C(0) = Var(Z(s))$$

The semi-variogram stops increasing beyond a certain distance. The variogram reaches a plateau which is called the "sill". This is also the variance of the random field $C(0)$.