Recovery and analysis of intra-site spatial data

Issues:

1. Locations of activities – what happened where and why?
2. Locations of social groups – who lived where and why?
    - -class, gender, ethnicity, etc.

Sources of data:

1. patterns in the location of features: ***site structure***
    - -buildings
    - -pits
    - -fences

2. patterns in the horizontal distribution of artifacts

Multiple spatial scales:

-meters x 1/10
-meters
-meters x 10
-meters x 100 …

-But which is right? (see O'Connell 1993)

Patterns in the horizontal distribution of artifacts

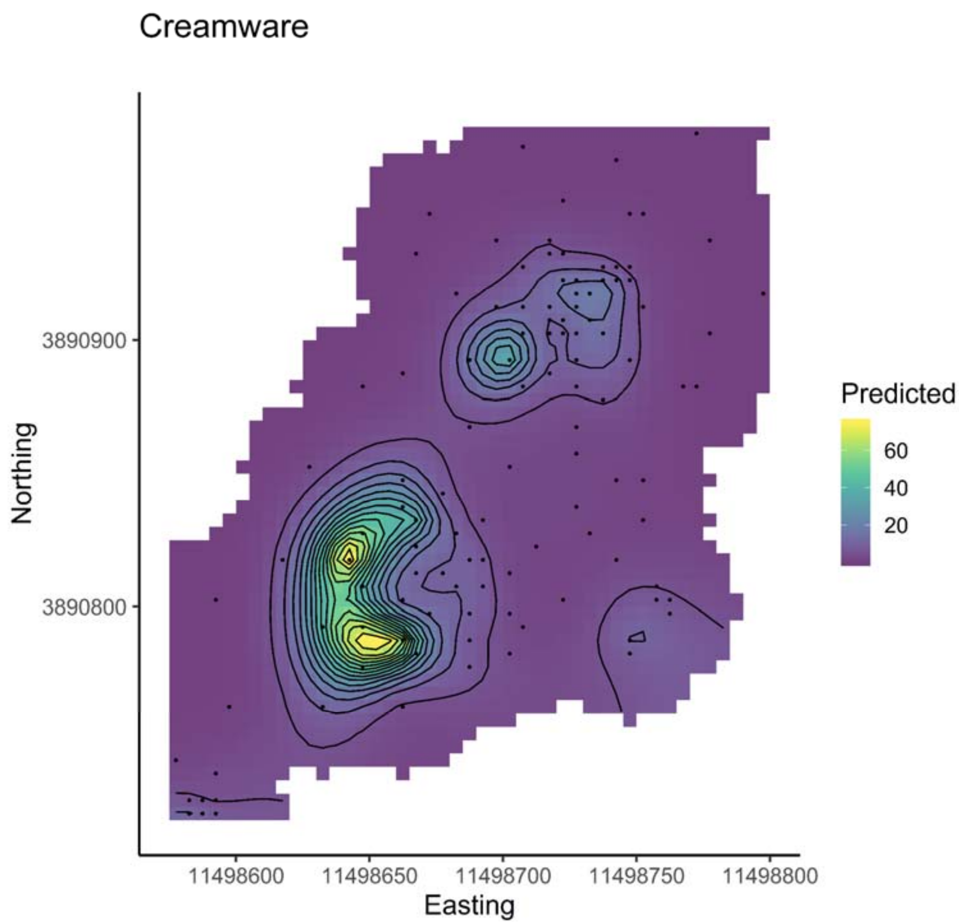
1. The spatial data recovery process

- Model how we collect spatial data.

Feedback:
Variogram

2. The analysis process

- Mapping raw data and statistical summaries of them.

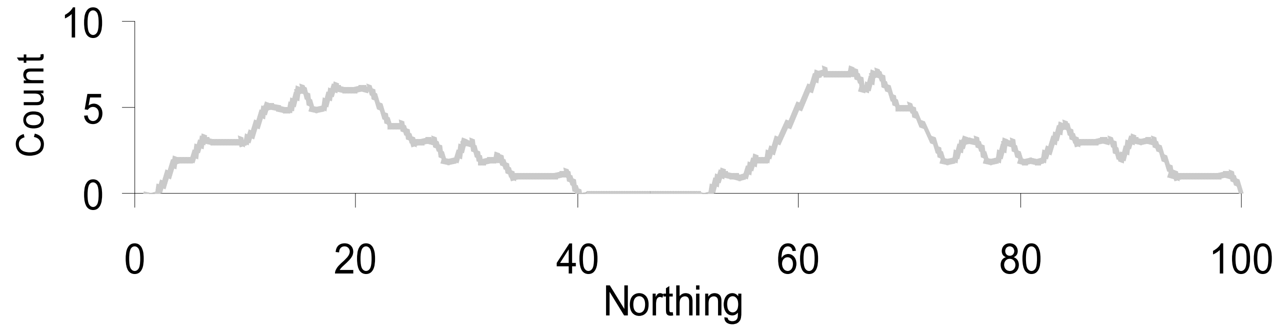


Recovery

The Point Process:

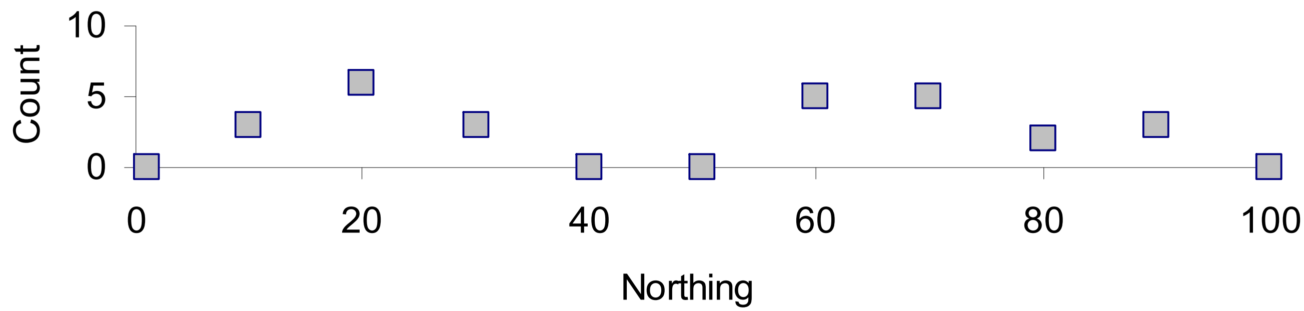
The Moving-Average Process:

(quadrat diameter=10)



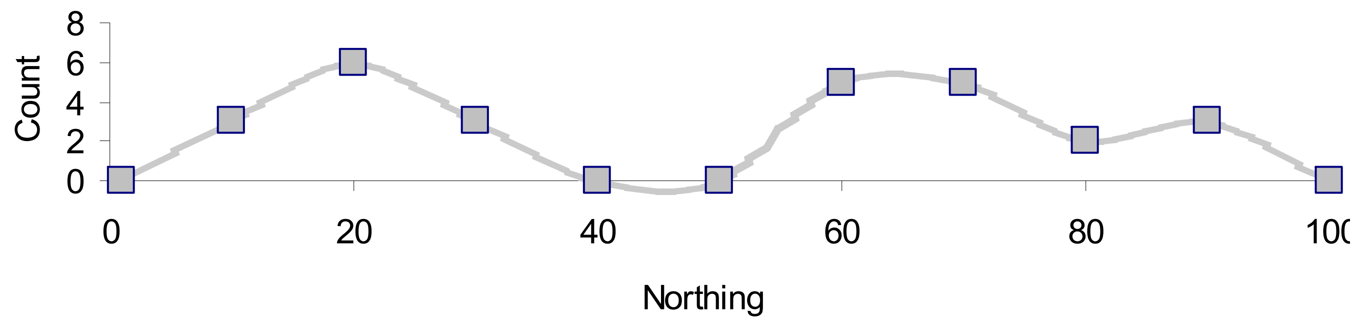
Sample the M-A Process

(quadrat spacing =10)

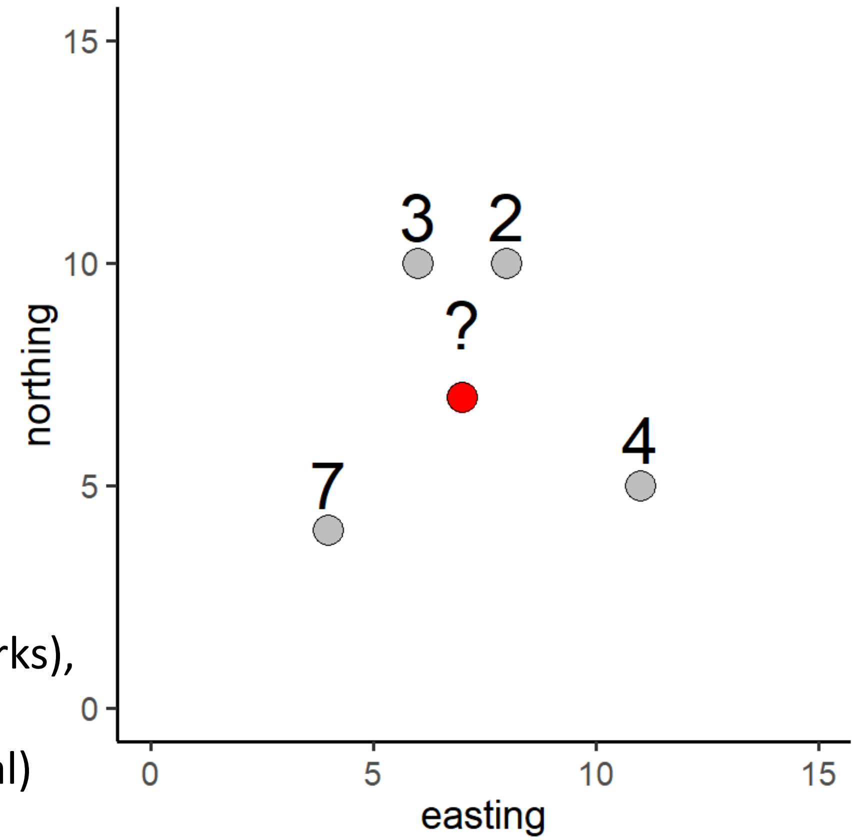


Analysis

Estimate the M-A Process from the sample:



Interpolation



Many methods…

1. Inverse distance weighting (IDW)

2. Kriging

3. Others
    - -TINs (triangulated irregular networks),
    - -splines (radial basis functions)
    - -polynomial regression (local, global)

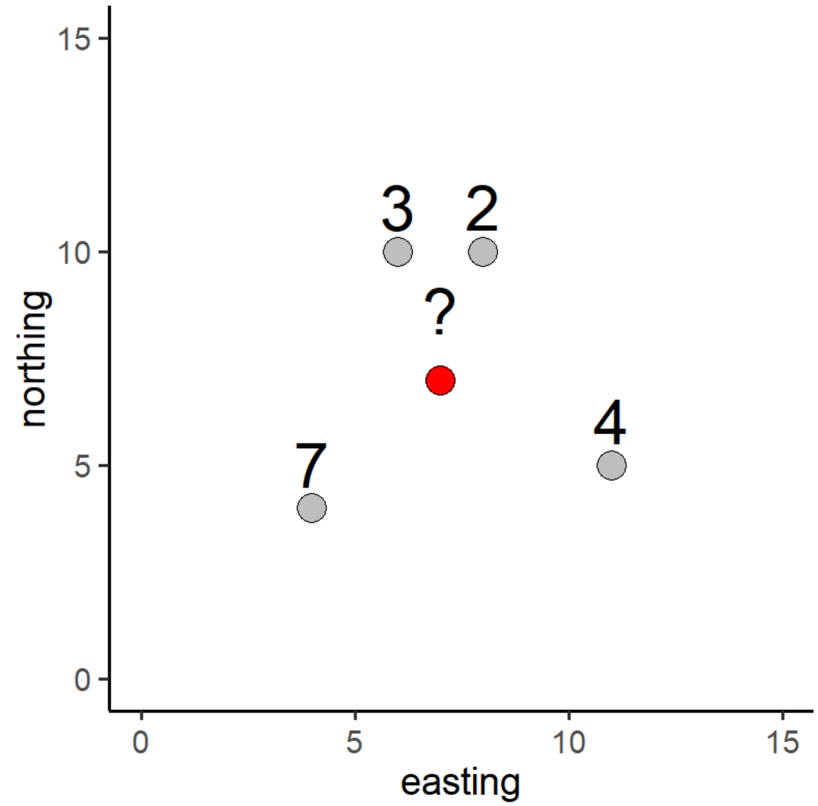
1. and 2. both make estimates of value of the z variable at an unsampled point in (x,y) space, as a ***weighted average of the values at nearby points, where z values are known***.

So….

$$\hat{z}_j = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{n} w_i}$$

Inverse Distance Weighting

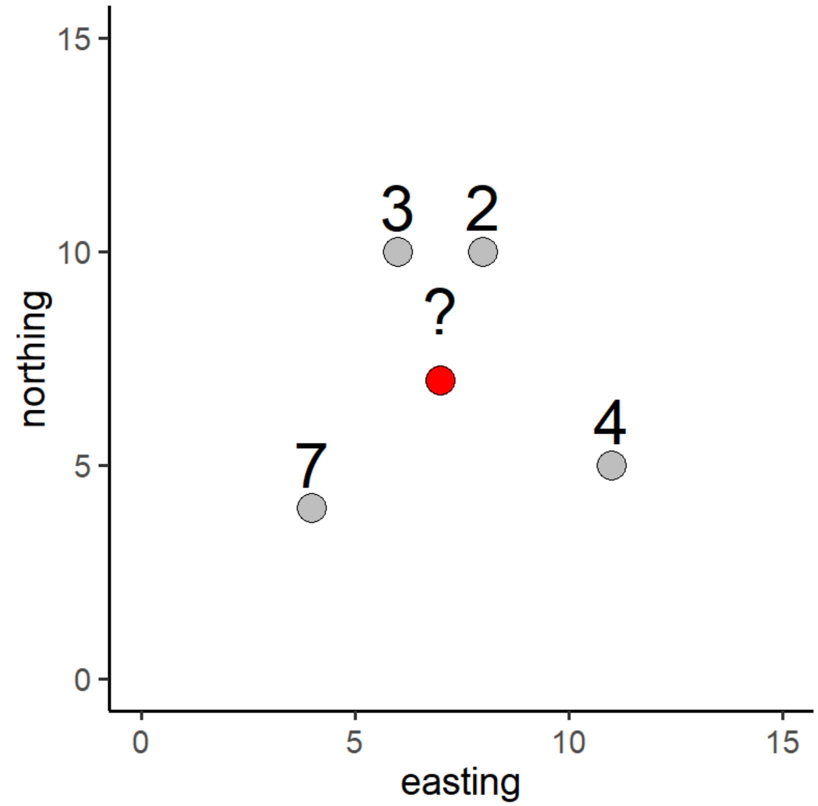
$$w_i = \frac{1}{d_{ij}^{\ p}}$$



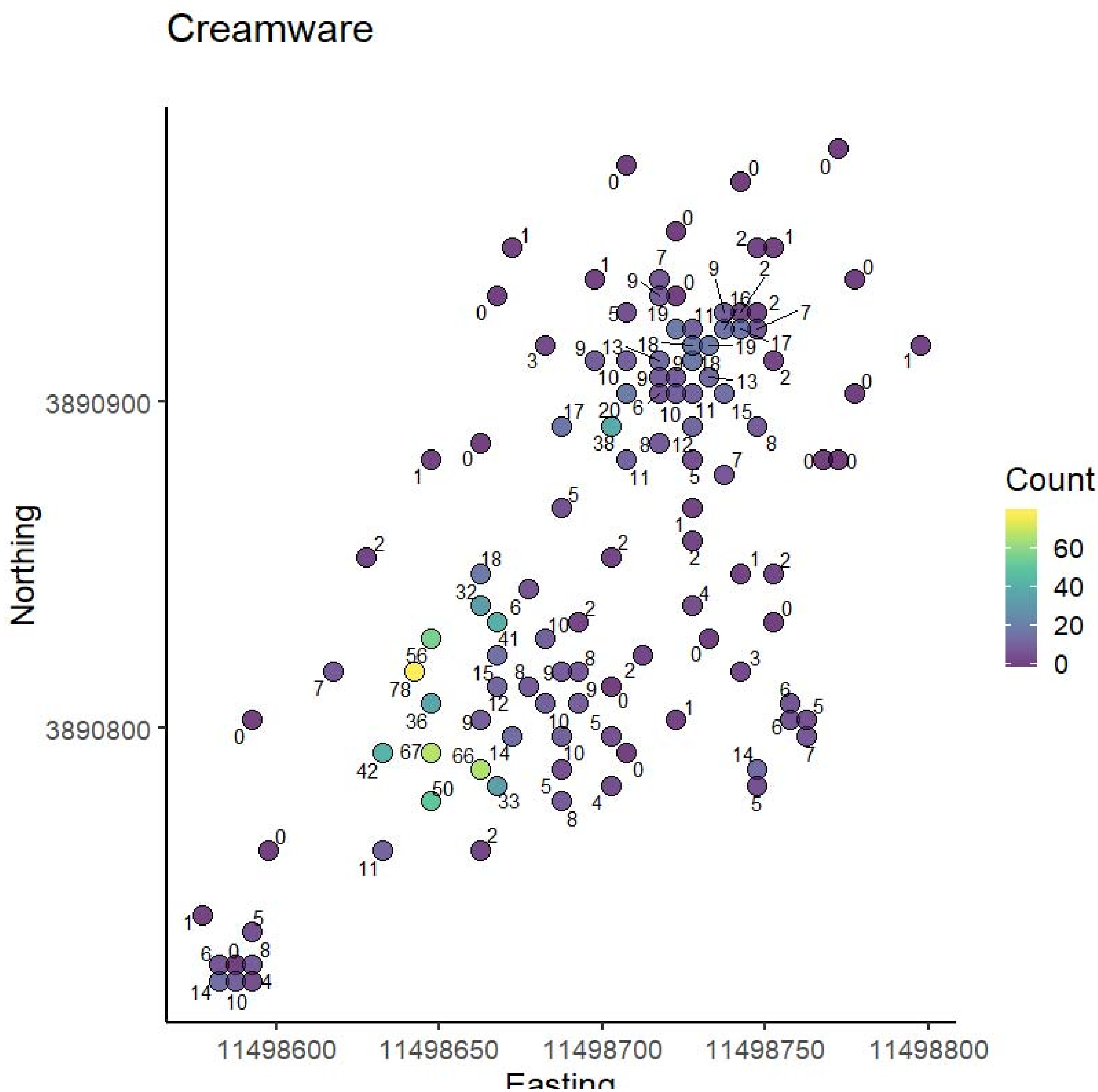
Point ID	northing	easting	z	$Distance = d_{ij}$	$w_i=1/d_{ij}$	$w_i{*}z$
1	*11*	*5*	*4*	*4.47*	*0.22*	*0.89*
2	*6*	*10*	*3*	*3.16*	*0.32*	*0.95*
3	*8*	*10*	*2*	*3.16*	*0.32*	*0.63*
4	*4*	*4*	*7*	*4.24*	*0.24*	*1.65*
Sum					*1.09*	*4.13*
5	*7*	*7*	*?*			

Inverse Distance Weighting

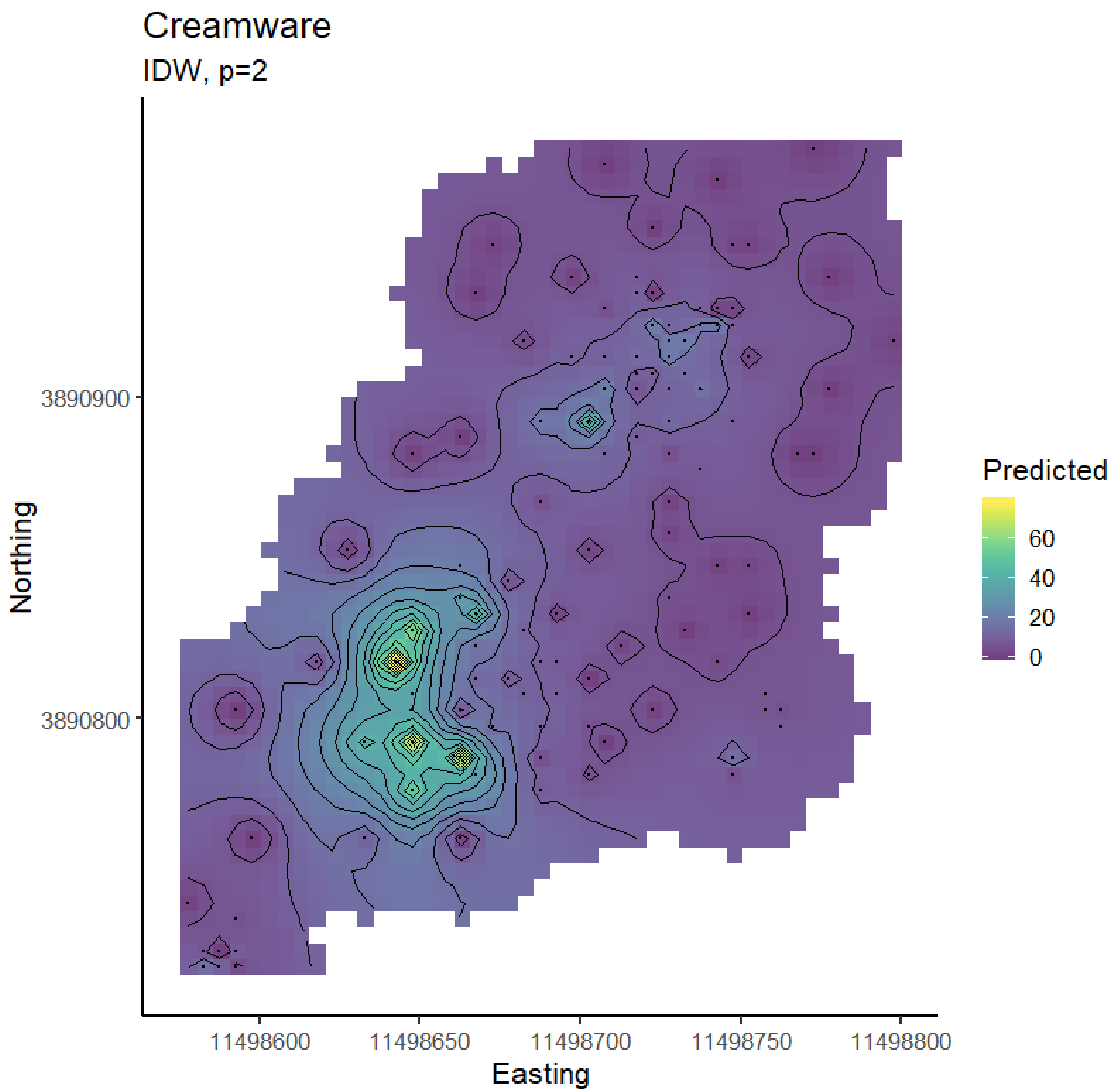
$$w_i = \frac{1}{{d_{ij}}^{p}}$$



Point ID	northing	easting	z	$Distance = d_{ij}$	$w_i=1/d_{ij}$	$w_i{*}z$
1	*11*	*5*	*4*	*4.47*	*0.22*	*0.89*
2	*6*	*10*	*3*	*3.16*	*0.32*	*0.95*
3	*8*	*10*	*2*	*3.16*	*0.32*	*0.63*
4	*4*	*4*	*7*	*4.24*	*0.24*	*1.65*
Sum					*1.09*	*4.13*
5	*7*	*7*	$4.13/1.09 = 3.8$			



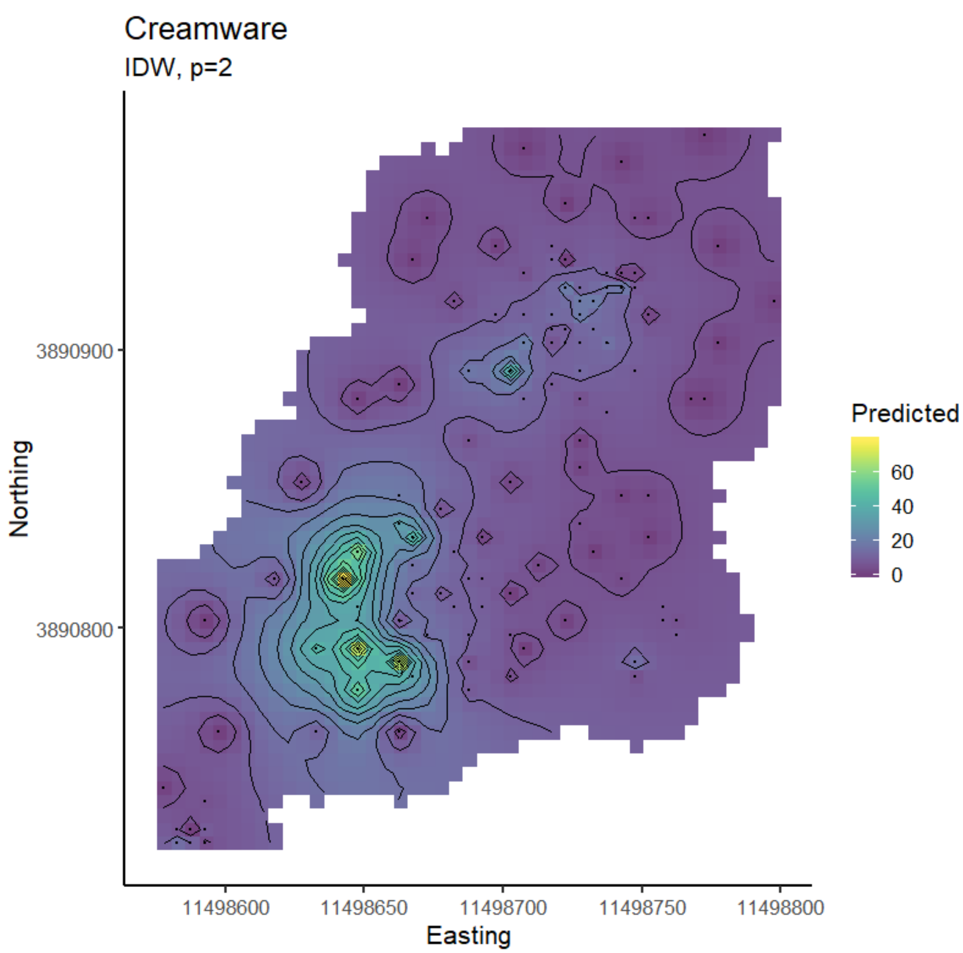
Creamware
Northing
3890900
3890800
Easting
11498600
11498650
11498700
11498750
11498800
Count
60
40
20
0

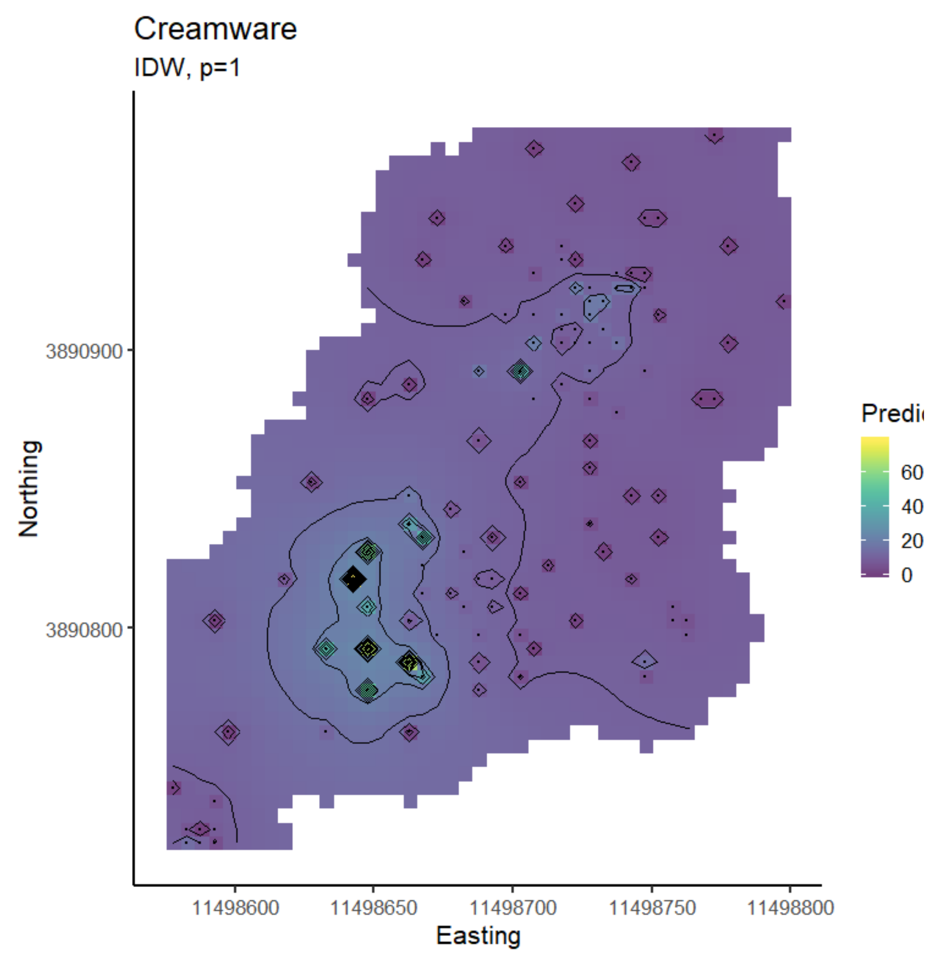


Creamware
IDW, p=2
Northing
3890900
3890800
11498600
11498650
11498700
11498750
11498800
Easting
Predicted
60
40
20
0

Pesky Questions about IDW

- what value for *p?*



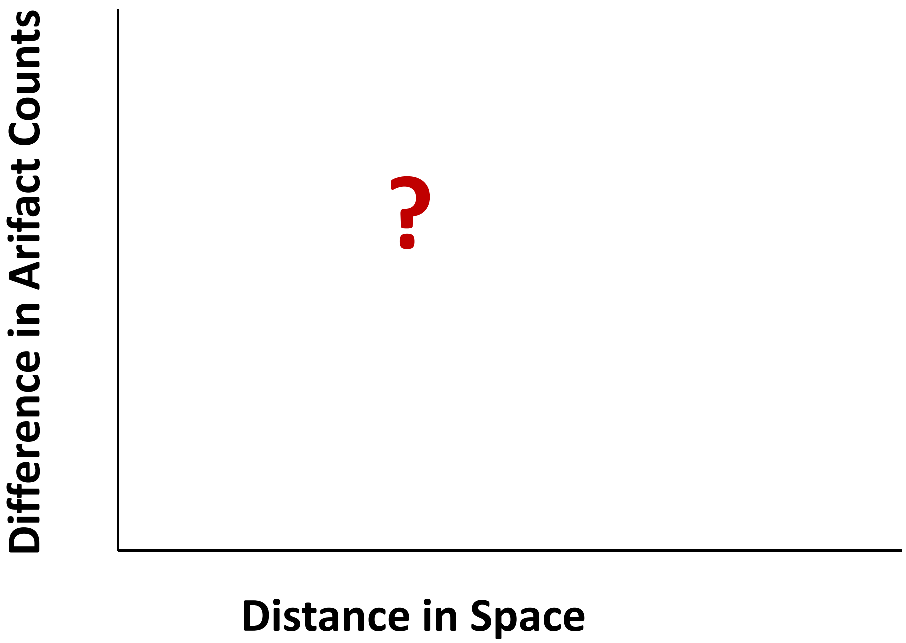


Doing Better than IDW

- *p* should depend on the manner in which ***differences*** between *z*-values increase with ***distances*** between *x,y* coordinates….

“Spatial autocorrelation”

To what extent do quadrats that are farther apart in 2-d space (*e.g*. Easting and Northing) tend to have variable values (*e.g. artifact counts*) that are more different.

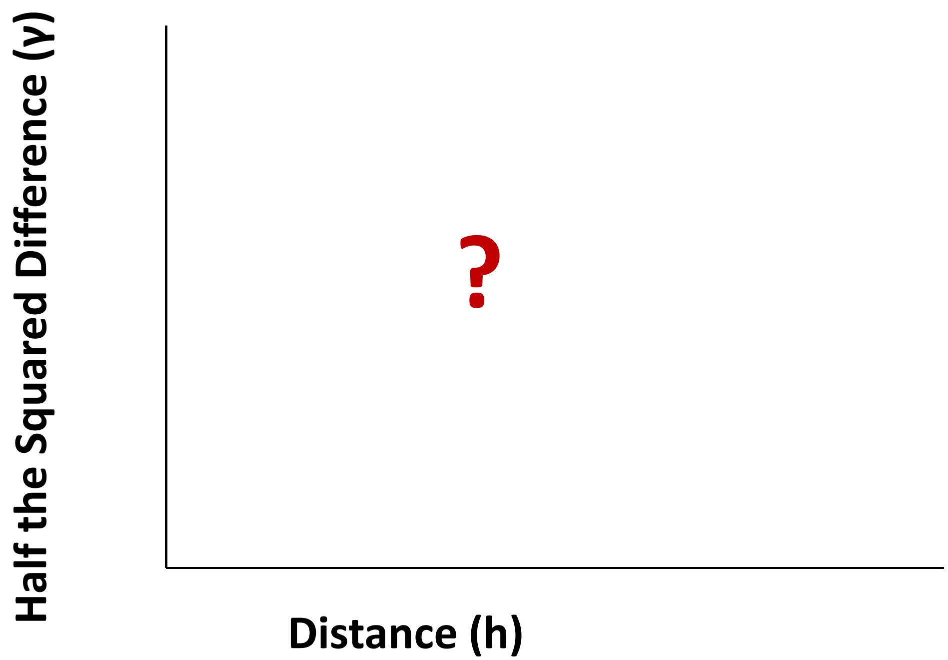


Kriging (after D.R Krige)

A weighted-averging interpolation method in which the ***weights depend on the spatial autocorrelation structure of the data***, AND that produces estimates of Z that are designed to minimize mean-squared prediction error.

Variogram

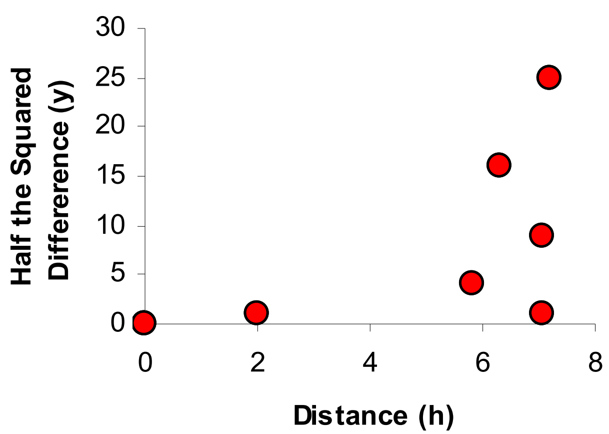
The graphical tool we use to measure the autocorrelation structure of spatial data.



The Variogram

Point	x	y	z
1	11	5	4
2	6	10	3
3	8	10	2
4	4	4	7

Distance From	To	*d*	Squared Differences
1	1	0	0
1	2	7.07	1
1	3	5.83	4
1	4	7.07	9
2	1	7.07	1
2	2	0	0
2	3	2	1
2	4	6.32	16
3	1	5.83	4
3	2	2	1
3	3	0	0
3	4	7.21	25
4	1	7.07	9
4	2	6.32	16
4	3	7.21	25
4	4	0	0

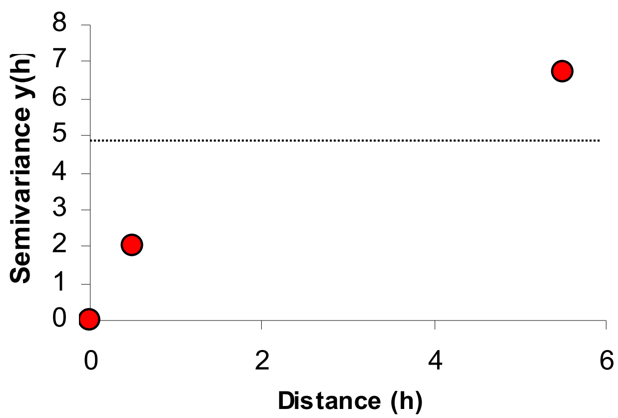


“**The variogram cloud**” – each graph point reoresent a difference-distance pair.

The Variogram

Point	x	y	z
1	11	5	4
2	6	10	3
3	8	10	2
4	4	4	7
	Variance		4.67

Distance From	To	h	Squared Differences	Distance Class	y(h)	mean(h)
1	1	0	0	0		
2	2	0	0	0		
3	3	0	0	0		
4	4	0	0	0	0	0
2	3	2	1	1-5		
3	2	2	1	1-5	0.5	2
1	3	5.831	4	5-10		
3	1	5.831	4	5-10		
2	4	6.325	16	5-10		
4	2	6.325	16	5-10		
1	2	7.071	1	5-10		
1	4	7.071	9	5-10		
2	1	7.071	1	5-10		
4	1	7.071	9	5-10		
3	4	7.211	25	5-10		
4	3	7.211	25	5-10	5.5	6.70175

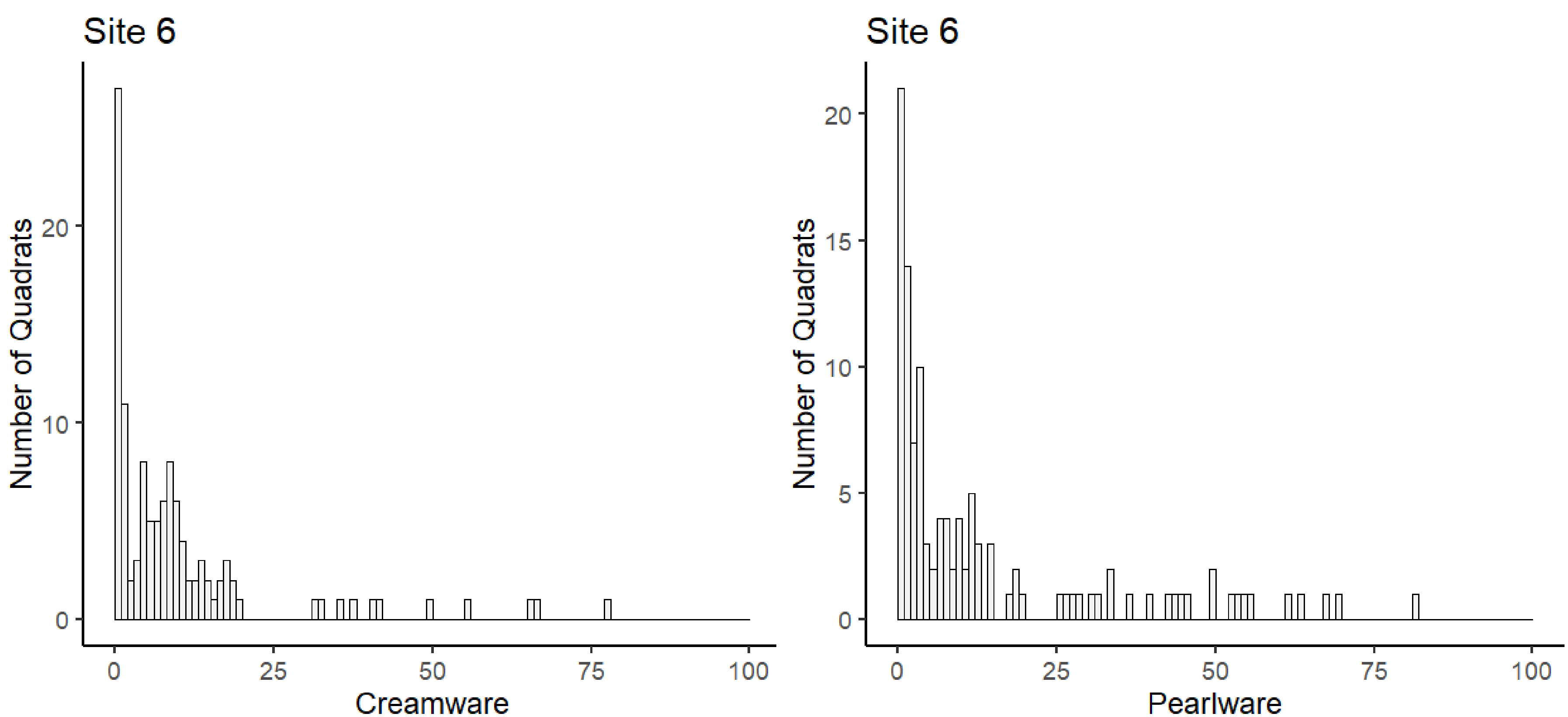


"**The variogram**" – each graph point represents the means of several difference-distance pairs.

The Variogram

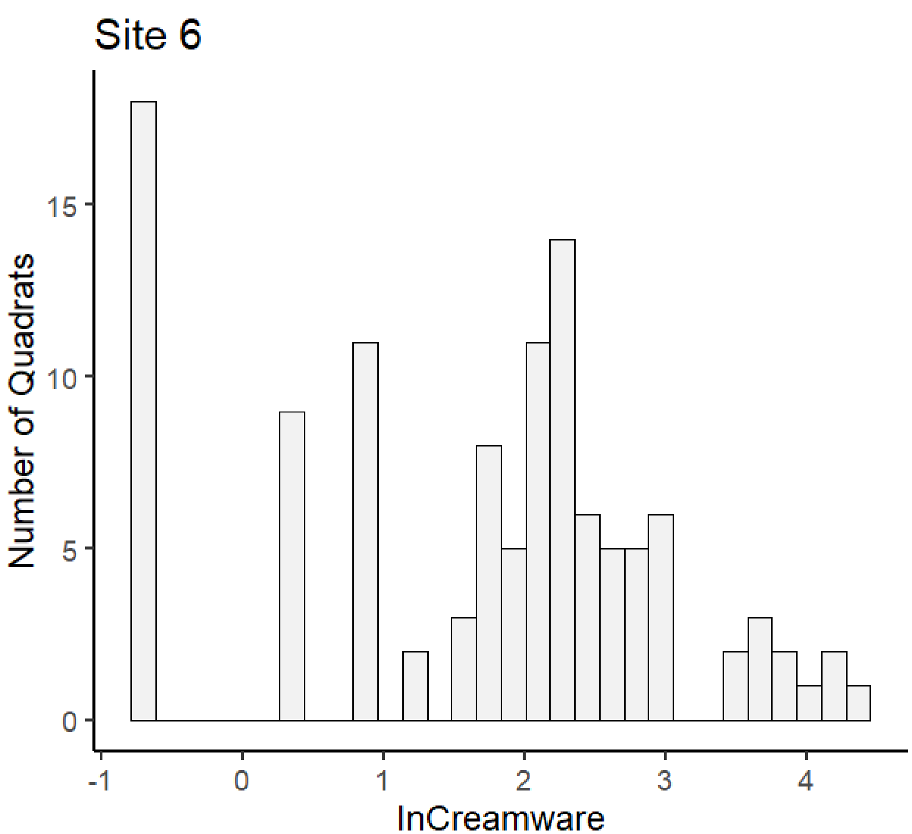
- The mathematical model behind the variogram and kriging assumes that the spatially distbuted variable has a normal of Gaussian distribution.

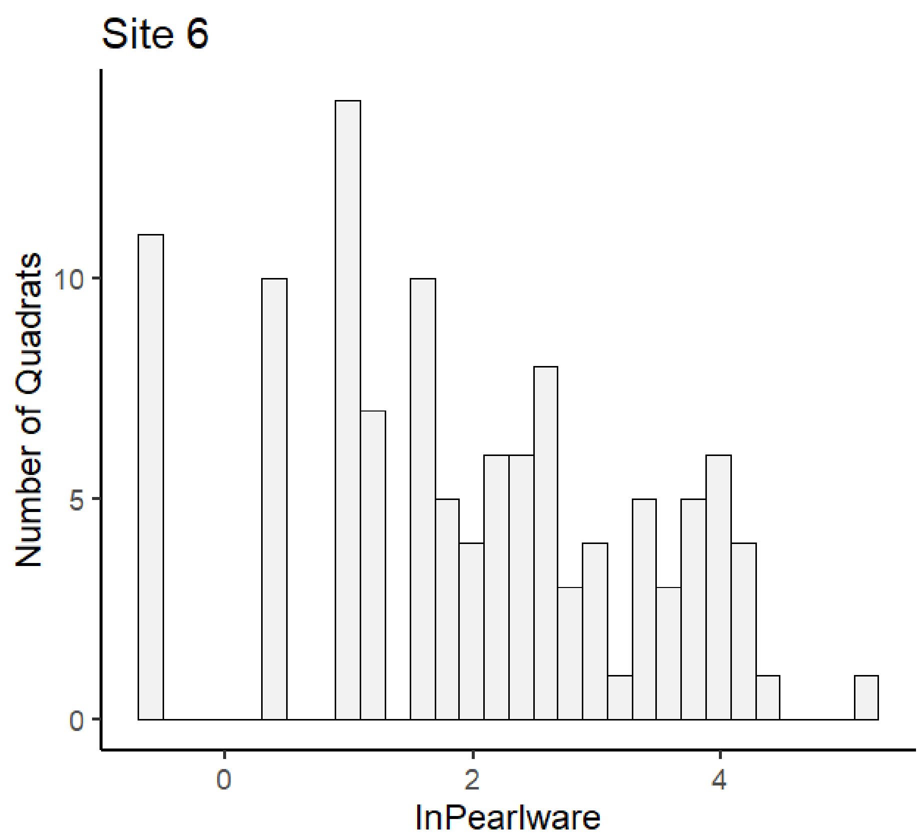
- But artifact counts always have long right tails…



The Variogram

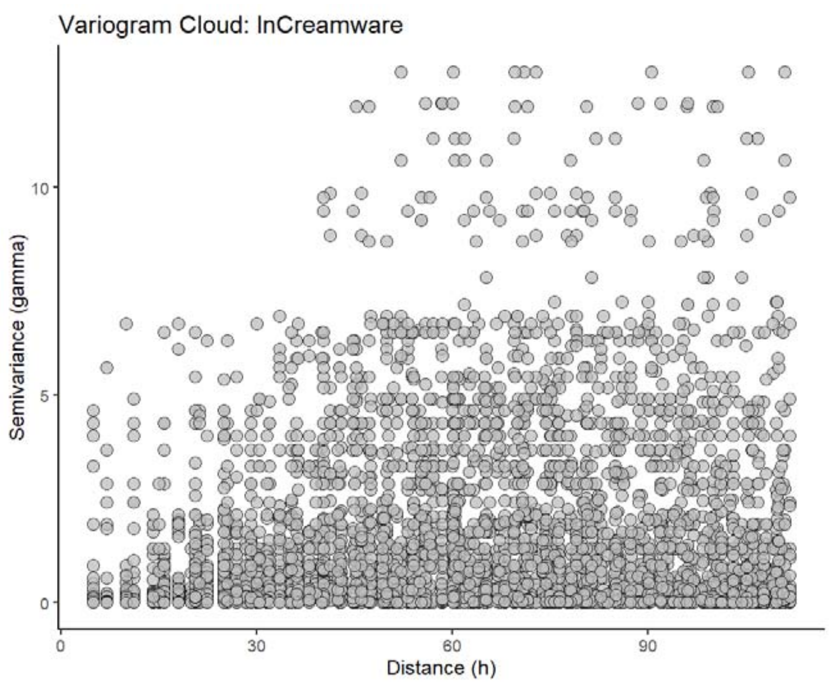
- Transforming the counts to a log scale helps.
- Because *ln(0)* is undefined, we take logs of "started counts"
    - *e.g. ln(Creamware +.5)*

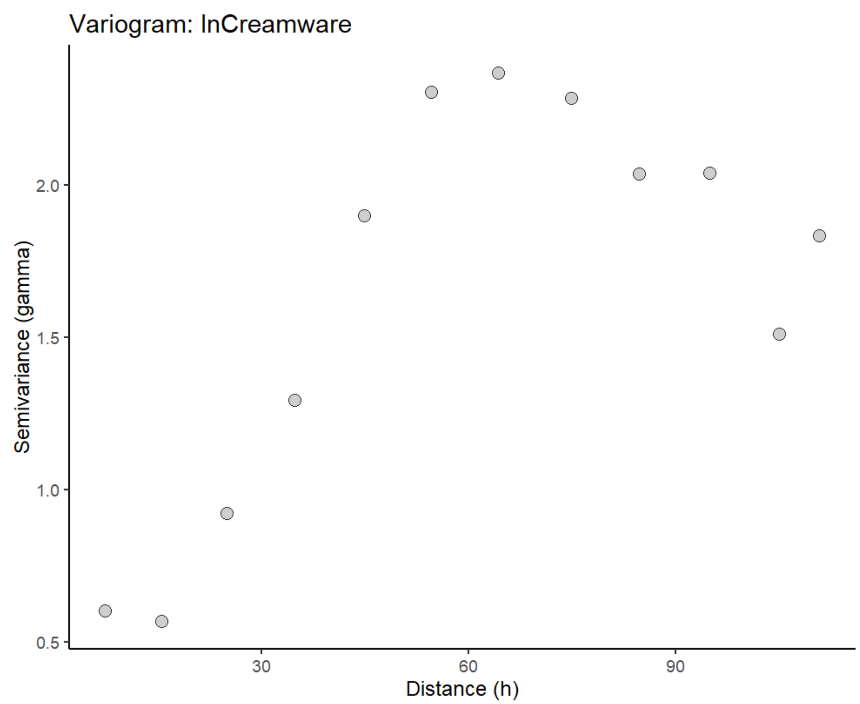




The Variogram

- Transforming the counts to a log scale helps.
- Because *ln(0)* is undefined, we take logs of "started counts"
    - *e.g. ln(Creamware +.5)*





The Variogram

Add up all the difference in pairs of z values

Square the differences in pairs of z values

$$\gamma_{\boldsymbol{h}} = \frac{1}{2n(\boldsymbol{h})}\sum_{i=1}^{n}(z(x_i, y_i) - z((x_i, y_i) + h))^2$$

The semivariance for distance class $\boldsymbol{h}$

$n(\boldsymbol{h})$ = the number of point pairs that fall in distance class $\boldsymbol{h}$

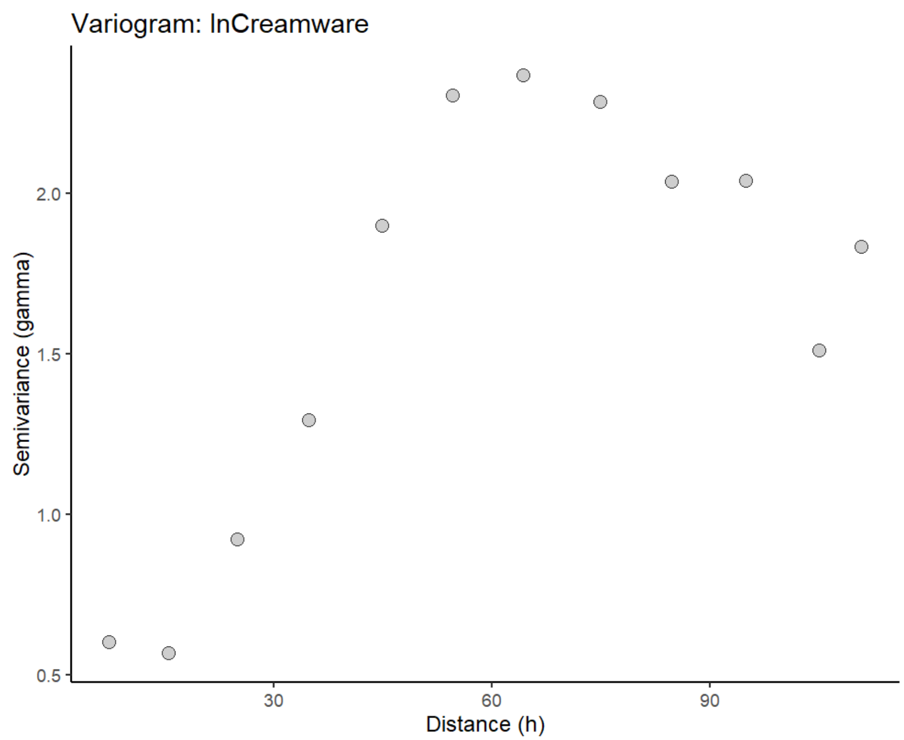
$z(x_i, y_i)$ = the value of z at the i'th point with coordinates x,y

$z(x_i, y_i)$+**h** = the value of z at the i'th point that is $\boldsymbol{h}$ *away from the point with* with coordinates x,y

(x, y) are 2-d spatial coordinates (easting, northing)

z is the variable value (artifact counts)

h is a distance and direction vector: "all the points that are a certain distance apart from the i'th *x,y* pair".



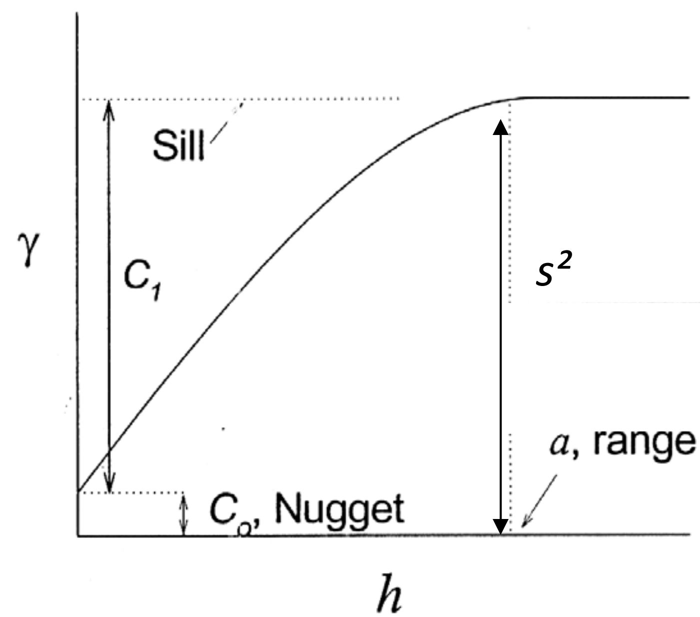
Variogram Lingo

- **Sill**: for larger values of *h* the variogram levels out, indicating that there no longer is any auto correlation between data points.

- *If the data are "well behaved" (Gaussian and stationary) the sill should be equal to the* ***variance (s**^{2}*) of the z values.*

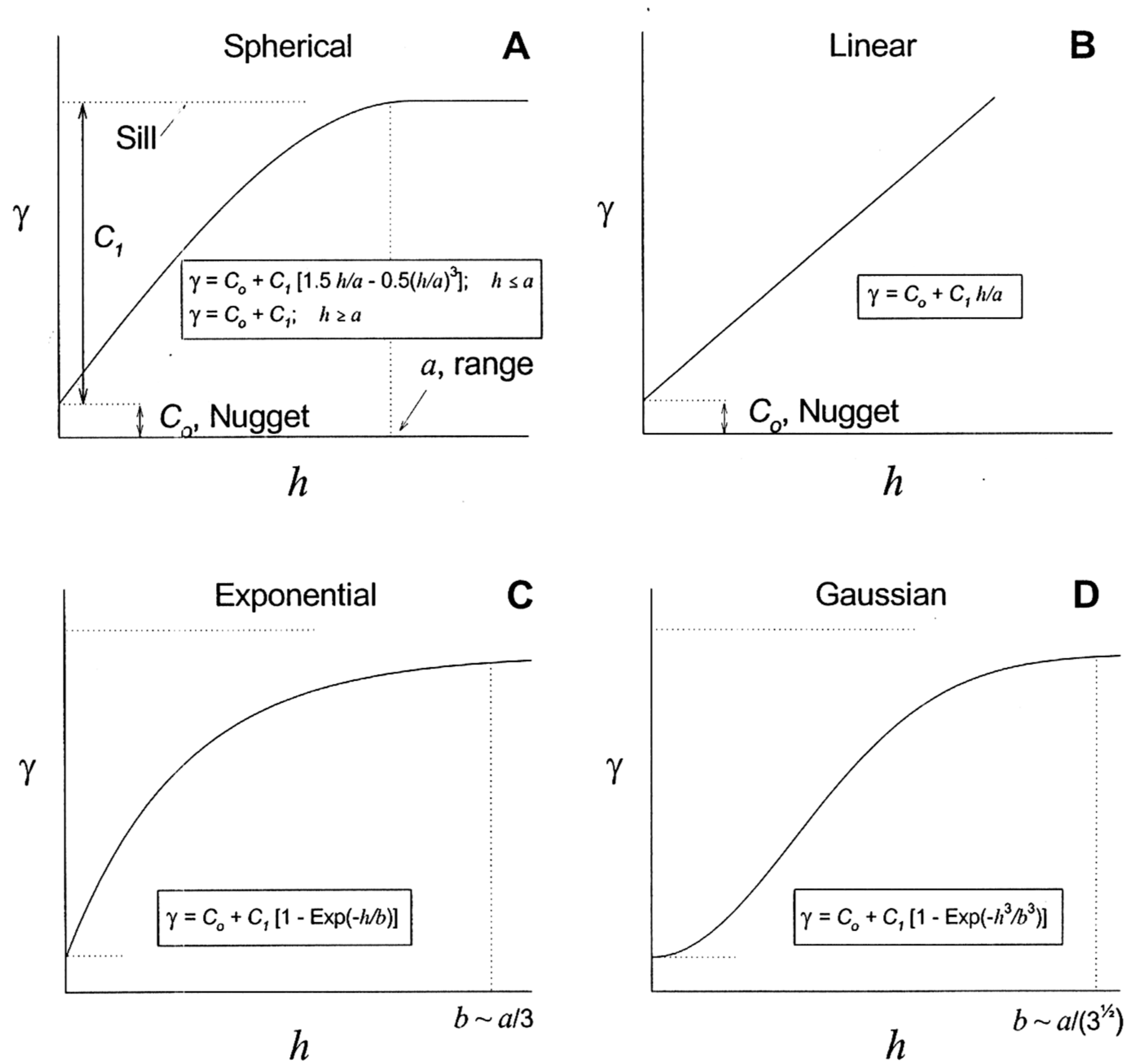
- **Range**: is the value of *h* where the sill occurs (or 95% of the value of the sill).
  This is the distance beyond which pairs of values are no longer autocorrelated.

- **Nugget variance**: a non-zero value for *gamma* when *h* = 0. Produced by various sources of unexplained error (e.g. measurement error).



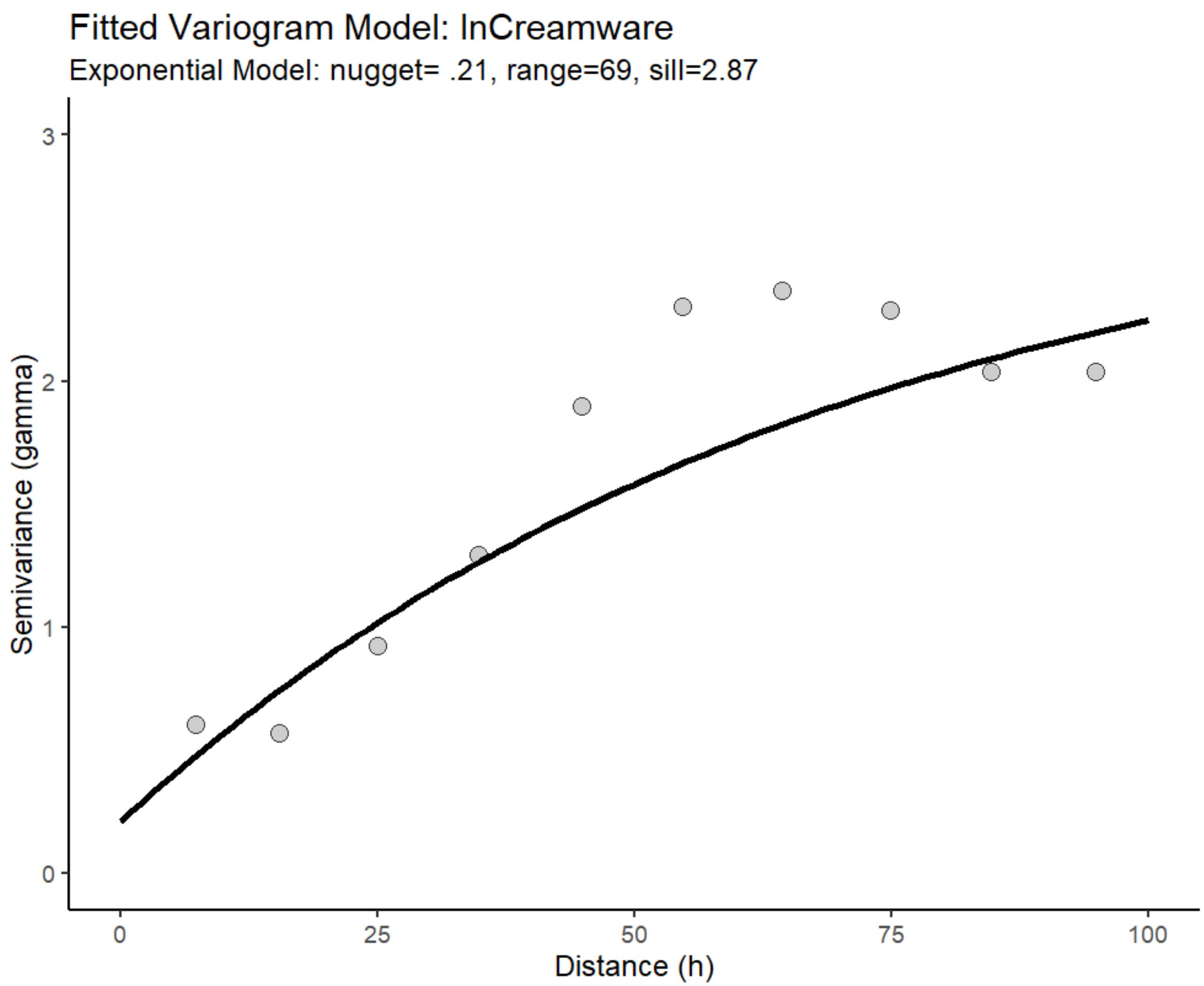
The Variogram

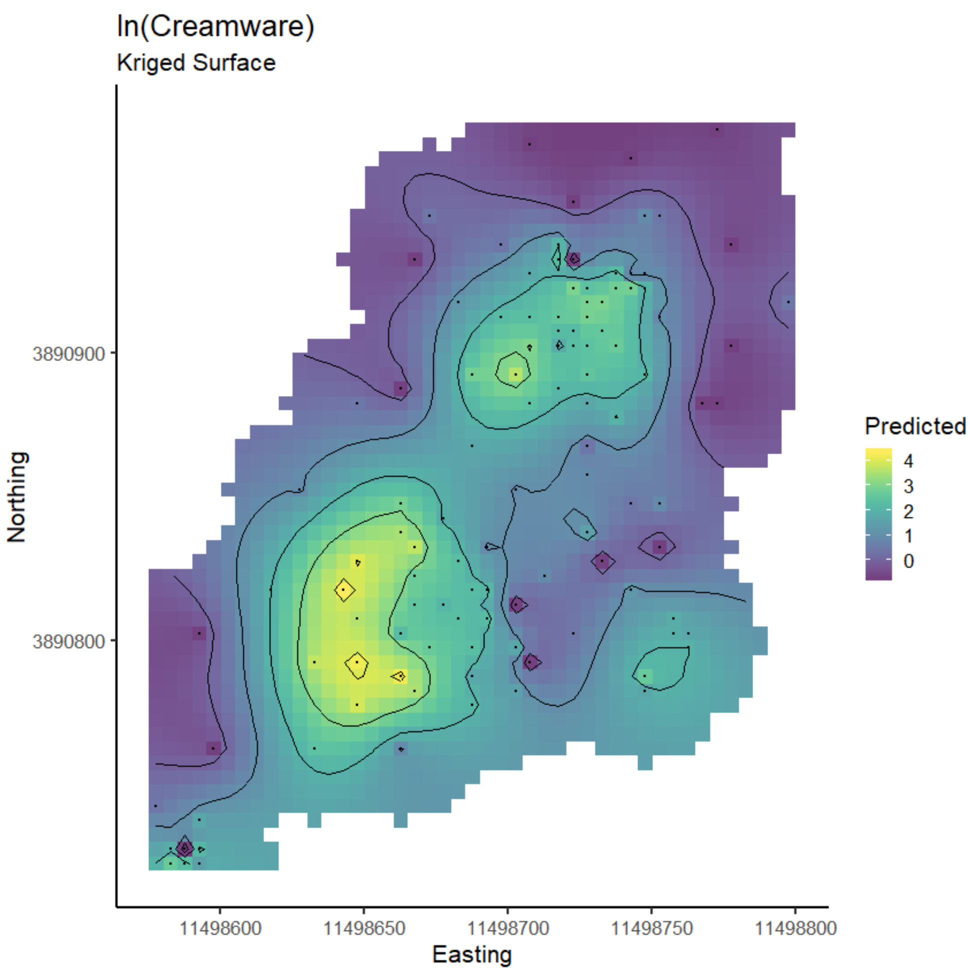
Variogram Models:
Differently shaped curves, defined by different equations.

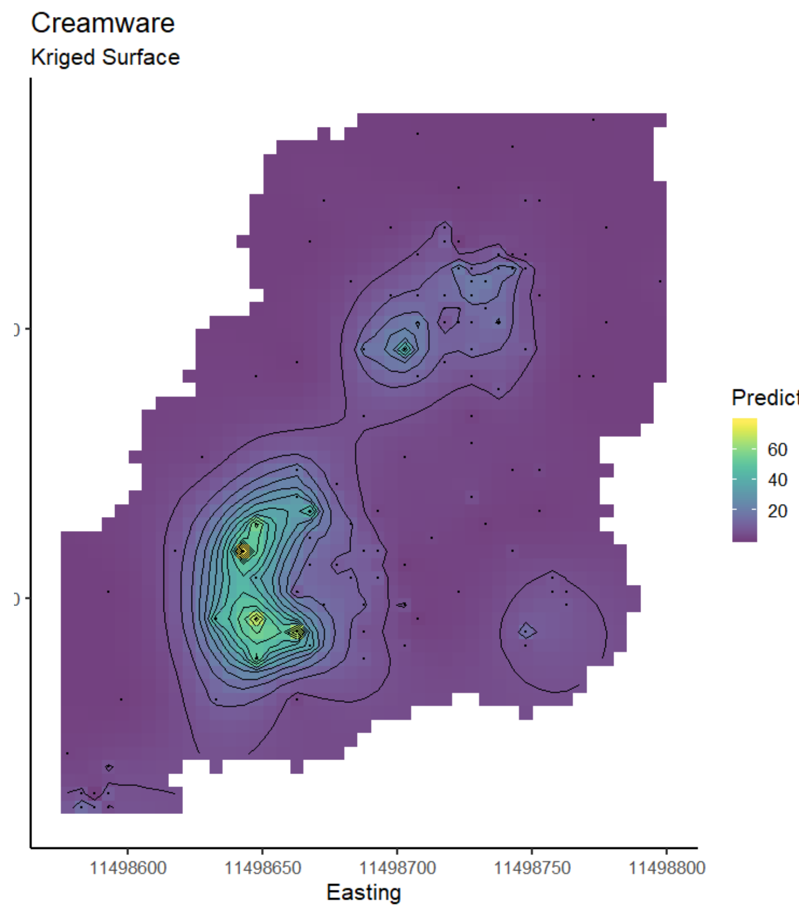


The Variogram

Variogram Models: Differently shaped curves, defined by different equations.




ln(Creamware)
Kriged Surface
Northing
3890900
3890800
11498600
11498650
11498700
11498750
11498800
Easting
Predicted
4
3
2
1
0


Creamware
Kriged Surface
11498600
11498650
11498700
11498750
11498800
Easting
60
40
20

The variogram is a useful spatial data analysis tool !!

You can use it during and excavation to see if your spatial sampling stratrgy is sufficient to capture spatial pattering

- Quadrat size (too small?)

- Quadrat spacing (too far part?)

- Given quadrat size and spacing is interpolation reasonable?