Lecture 6: Point Level Models - Variograms & EDA

Class Intro

Intro Questions

- What is a variogram?
- How is a variogram useful?
- For Today:
  - How is EDA used for point referenced data?

Variogram Creation

Variogram Creation: How?

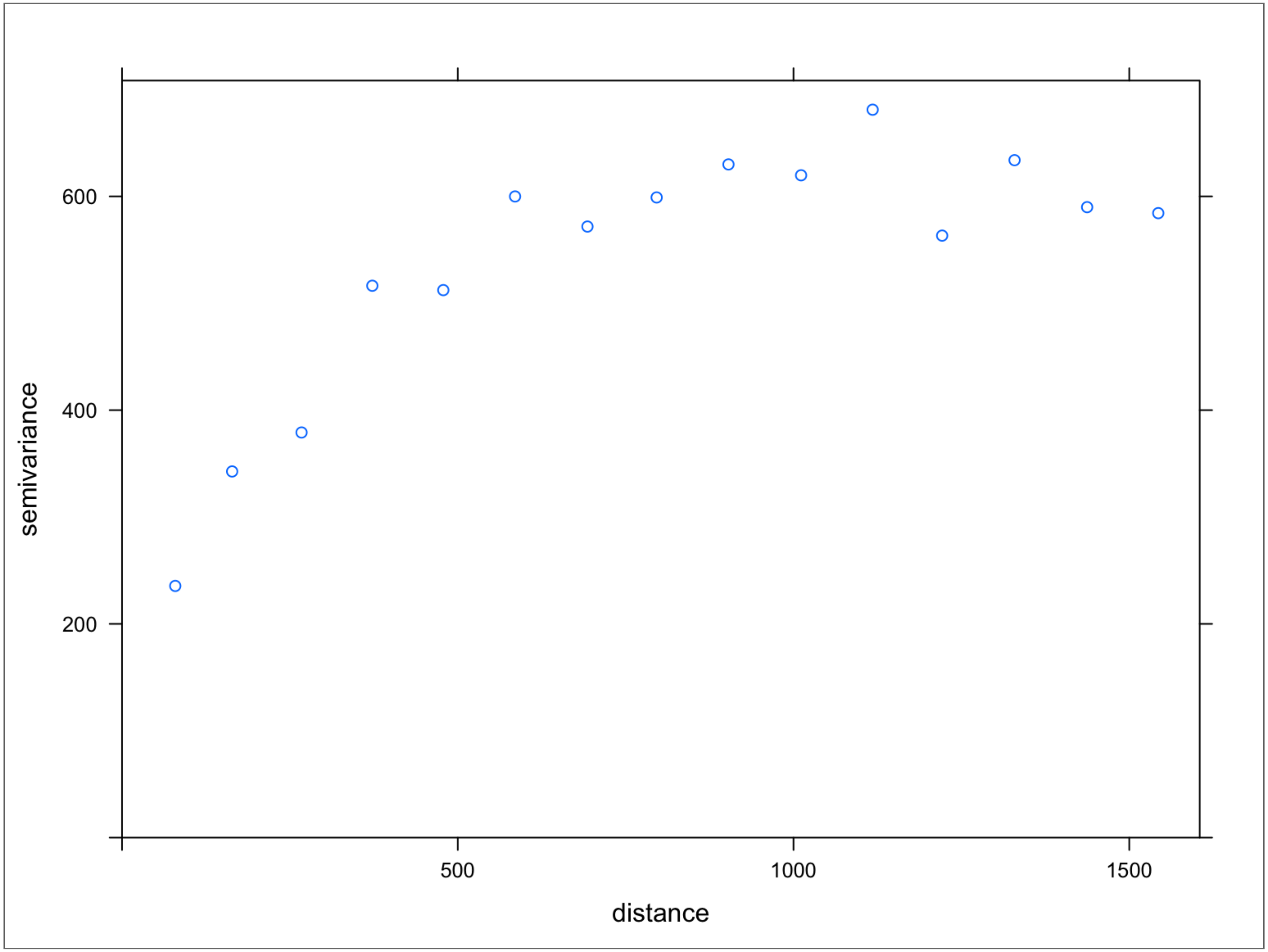
x	y	copper
181072	333611	85
181025	333558	81
181165	333537	68
181298	333484	81
181307	333330	48
181390	333260	61
181165	333370	31
181027	333363	29
181060	333231	37
181232	333168	24
181191	333115	25
181032	333031	25
180874	333339	93

x	y	copper
180969	333252	31
181011	333161	27

Variogram Creation: Steps

1. Calculate distances between sampling locations
2. Choose grid for distance calculations
3. Calculate empirical semivariogram
   $$\hat{\gamma}(d_k) = \frac{1}{2N(d_k)} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_i, \in N(d_k)} \left[\boldsymbol{Y}(\boldsymbol{s}_i) - \boldsymbol{Y}(\boldsymbol{s}_j)\right]^2,$$
   where
   $N(d_k) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : ||\boldsymbol{s}_i - \boldsymbol{s}_j|| \in I_k\}$
   and I_k is the k^{th} interval.
4. Plot the semivariogram

Variogram Creation: R function



Variogram Creation: How - Step 1

Calculate Distances between sampling locations

```
dist(meuse.small)
```

```
##              1          2          3          4          5          6
## 2     70.95069
## 3    120.05832  142.16188
## 4    259.27013  282.85155  143.76022
## 5    368.17795  364.13871  251.81938  157.75297
## 6    474.23728  471.62379  356.93557  242.98148  109.35264
## 7    263.90529  239.67478  171.04970  182.16751  148.50253  252.23997
## 8    258.19566  201.82418  225.47949  301.30715  282.57742  378.68457
## 9    383.20752  331.79813  324.99538  350.12712  266.32874  332.14003
## 10   474.94210  445.19434  377.60561  327.81245  180.12496  186.53954
## 11   513.59225  476.38325  424.98118  388.26022  245.37726  248.84131
## 12   584.46557  530.01321  524.95143  528.30010  406.88450  426.49853
## 13   336.52934  266.28181  352.85975  448.26889  435.42508  522.99235
## 14   377.36720  315.07459  347.86492  405.66612  347.29958  422.14334
## 15   457.80454  400.90024  408.37850  435.44690  341.49378  393.18952
## 16   437.93835  367.95924  444.10584  525.06476  485.82816  560.73256
## 17   593.84426  524.22228  590.87139  656.25757  589.65074  646.24608
## 18   742.60420  673.12777  735.57257  791.89898  710.19293  753.27551
## 19   885.30277  815.88296  876.18548  926.73729  835.97488  869.40094
## 20  1041.44371  972.26385 1030.40235 1075.16231  977.67172 1002.09281
## 21  1000.66628  932.72343  980.30352 1016.51168  910.56795  929.66230
## 22   968.07025  901.35343  941.28104  971.20595  859.96919  875.28338
## 23  1017.21827  950.99579  987.79502 1013.58374  898.51767  909.11385
## 24   693.45368  648.32245  613.69699  580.30595  433.95507  412.02306
## 25   793.75689  747.08366  715.46069  680.25510  532.10243  501.24146
```

```
dist.mat <- dist(meuse.small %>% select(x,y))
```

Variogram Creation: How - Step 2

Choose grid for distance calculations

```
cutoff <- max(dist.mat) / 3 # default maximum distance
num.bins <-  15
bin.width <- cutoff / 15
```

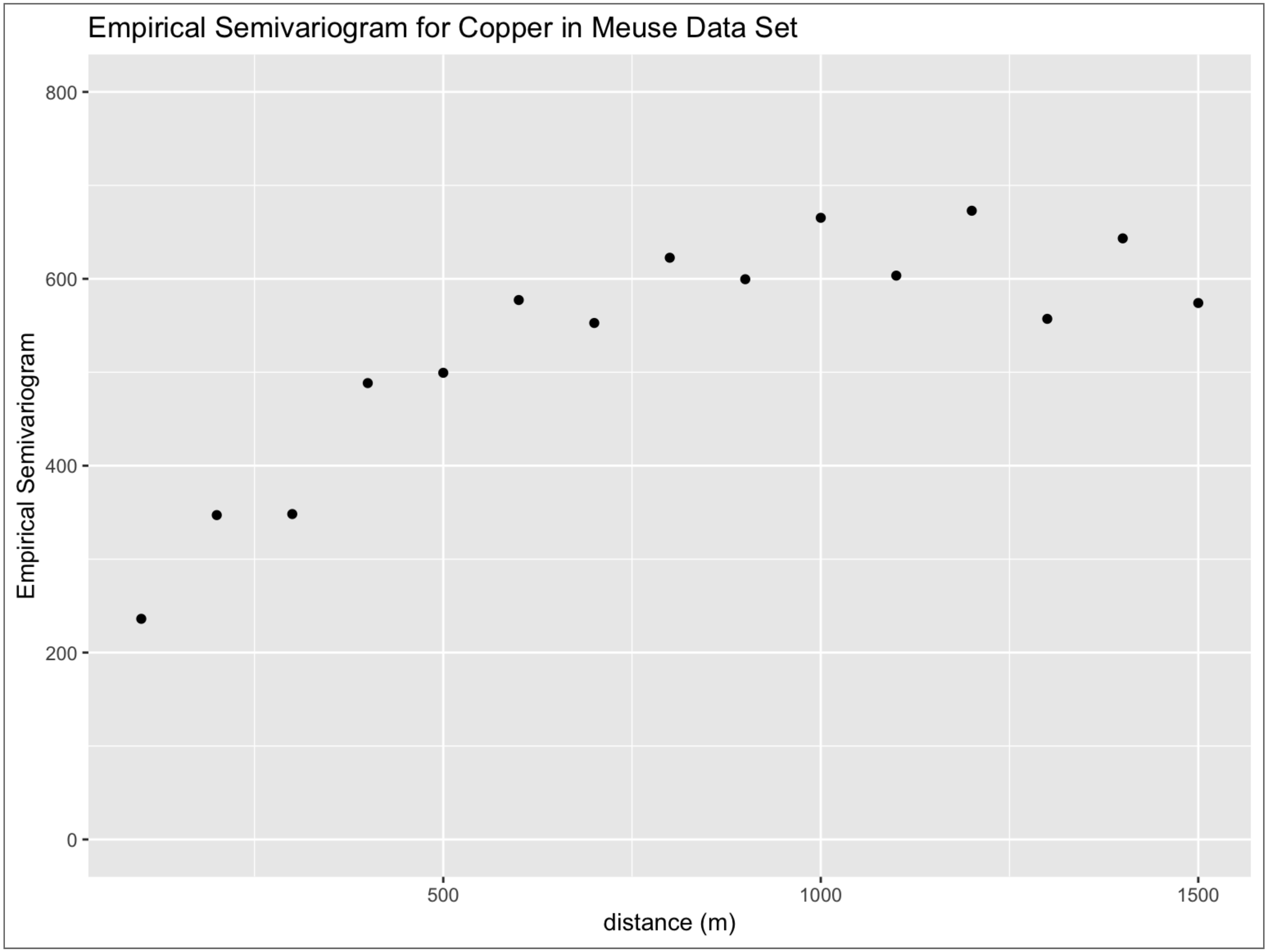
Variogram Creation: How - Step 3

Calculate empirical semivariogram

```
## # A tibble: 15 x 2
##      bin emp.sv
##    <dbl>  <dbl>
##  1     1   236.
##  2     2   347.
##  3     3   348.
##  4     4   488.
##  5     5   499.
##  6     6   577.
##  7     7   553.
##  8     8   623.
##  9     9   600.
## 10    10   665.
## 11    11   603.
## 12    12   673.
## 13    13   557.
## 14    14   643.
## 15    15   574.
```

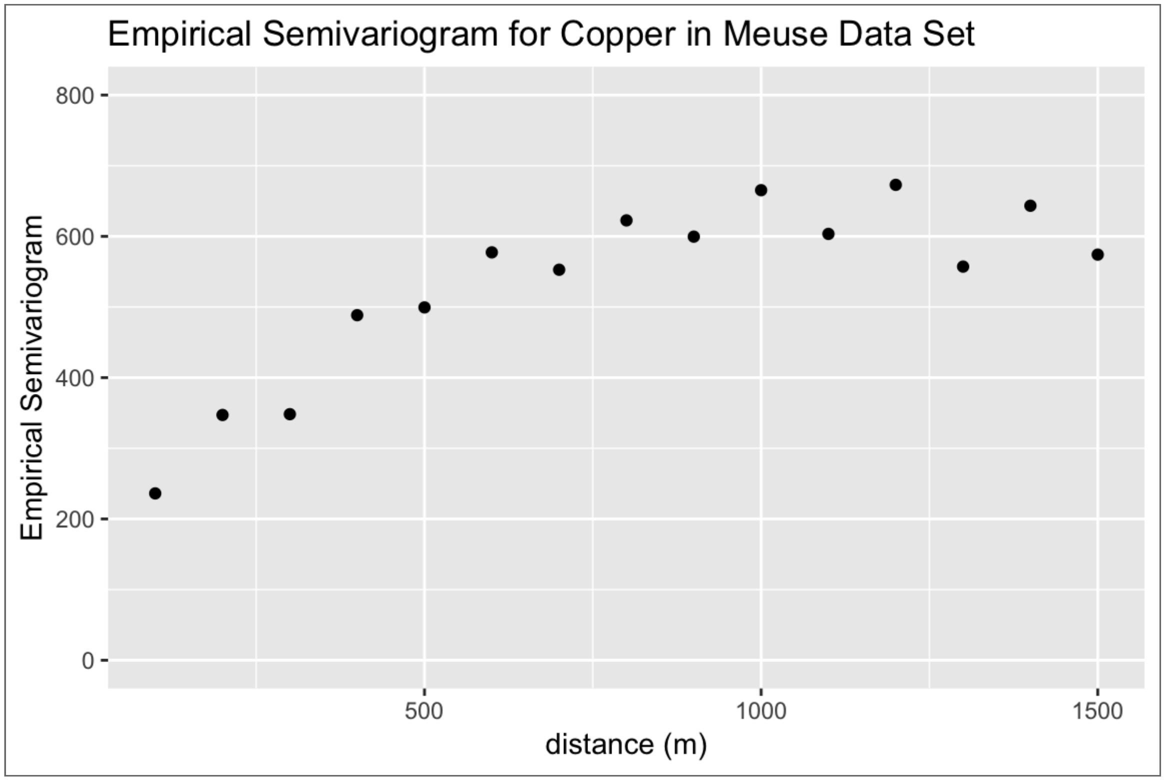
Variogram Creation: How - Step 4

Plot empirical semivariogram


Empirical Semivariogram for Copper in Meuse Data Set
Empirical Semivariogram
800
600
400
200
0
500
1000
1500
distance (m)

Variogram Fitting

Now given this empirical semivariogram, how to we choose a semivariogram (and associated covariance structure) and estimate the parameters in that function??



Variogram Fitting, cont..

- Empirical semivariograms can be computed and plotted, but they are subject to researcher choice: such as bin width and the number of points to display.
- The “data points” here is really a function of the observed random variable, but do not have an associated likelihood.
- Variogram fitting is very much an art and not a science.
- Eventually maximum likelihood or Bayesian methods will be used to estimate parameters in the covariance model directly.

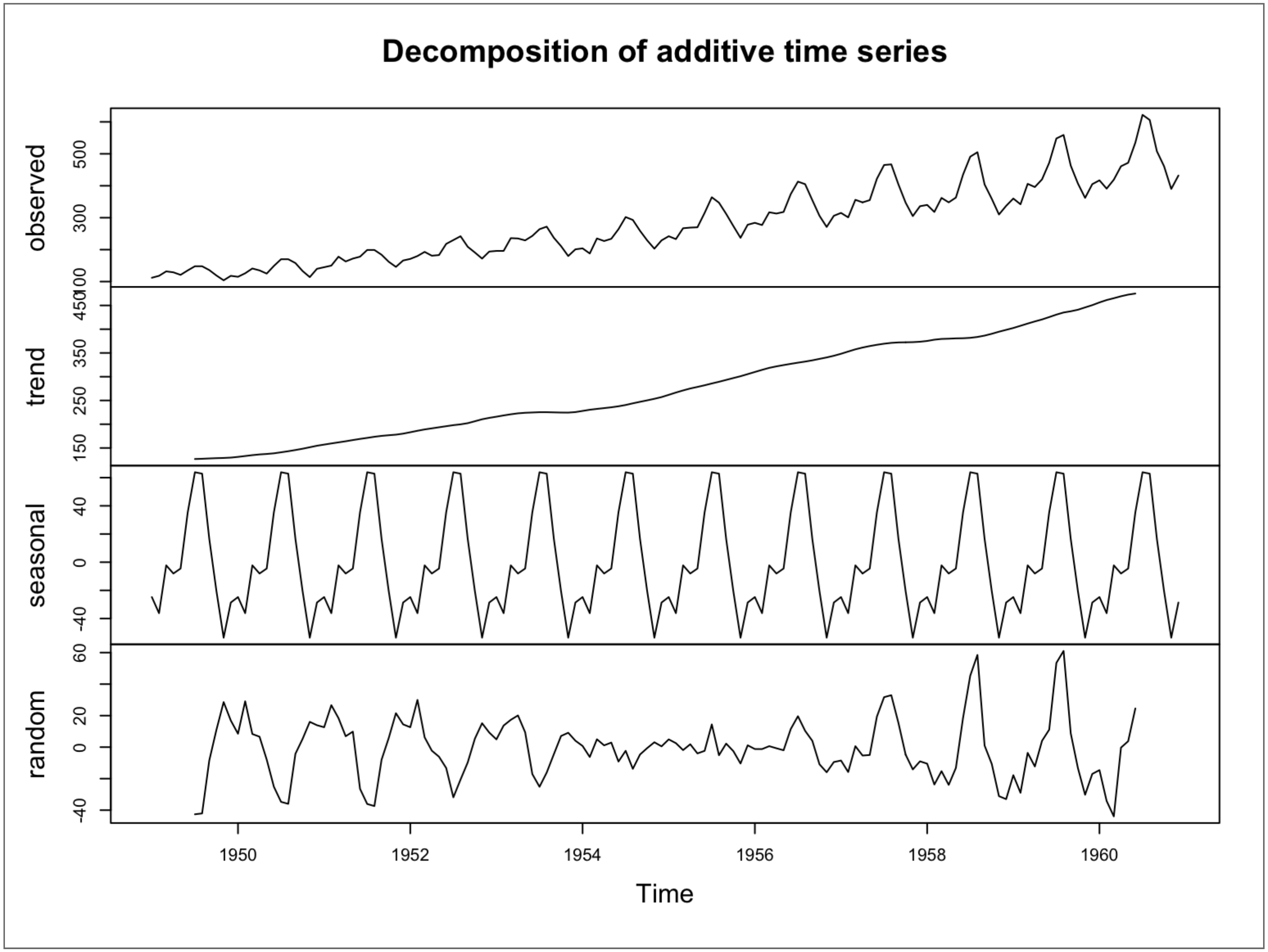
Exploratory Data Analysis

EDA Overview

- Exploratory Data Analysis (EDA) is commonly used to explore and visualize data sets.
- EDA is not a formal analysis, but can inform modeling decisions.
- What are we interested in learning about with spatial data?

Data Decomposition: Time Series

- In time series analysis, the first step in the EDA process was to decompose the observed data into a trend, seasonal cycles, and a random component.


Decomposition of additive time series
observed
500
300
100
trend
450
350
250
150
seasonal
40
0
-40
random
60
20
0
-40
1950
1952
1954
1956
1958
1960
Time

Data Decomposition: Spatial Data

- Similarly spatial data will be decomposed into the mean surface and the error surface.
- For example, elevation and distance from major bodies of water would be part of the mean surface for temperature.
- The mean surface is focused on the global, or first-order, behavior.
- The error surface captures local fluctuations, or second-order, behavior.

Response Surface vs. Spatial Surface

- Spatial structure in the response surface and spatial structure in the error surface are not one-and-the-same.
- $E[(Y(\boldsymbol{s}) - \mu)(Y(\boldsymbol{s}') - \mu)]$ vs. $E[(Y(\boldsymbol{s}) - \mu(\boldsymbol{s}))(Y(\boldsymbol{s}') - \mu(\boldsymbol{s}))]$
- There are stationarity implications for considering the residual surface.
- Data sets contain two general types of useful information: spatial coordinates and covariates.
- Regression models will be used to build the mean surface.

Spatial EDA Overview

1. Map of locations
2. Histrogram or other distributional figure
3. 3D scatterplot
4. General Regression EDA
5. Variograms and variogram clouds
6. Anistopic diagnostics

Scallops Data Example

1. Map of Locations

+
−

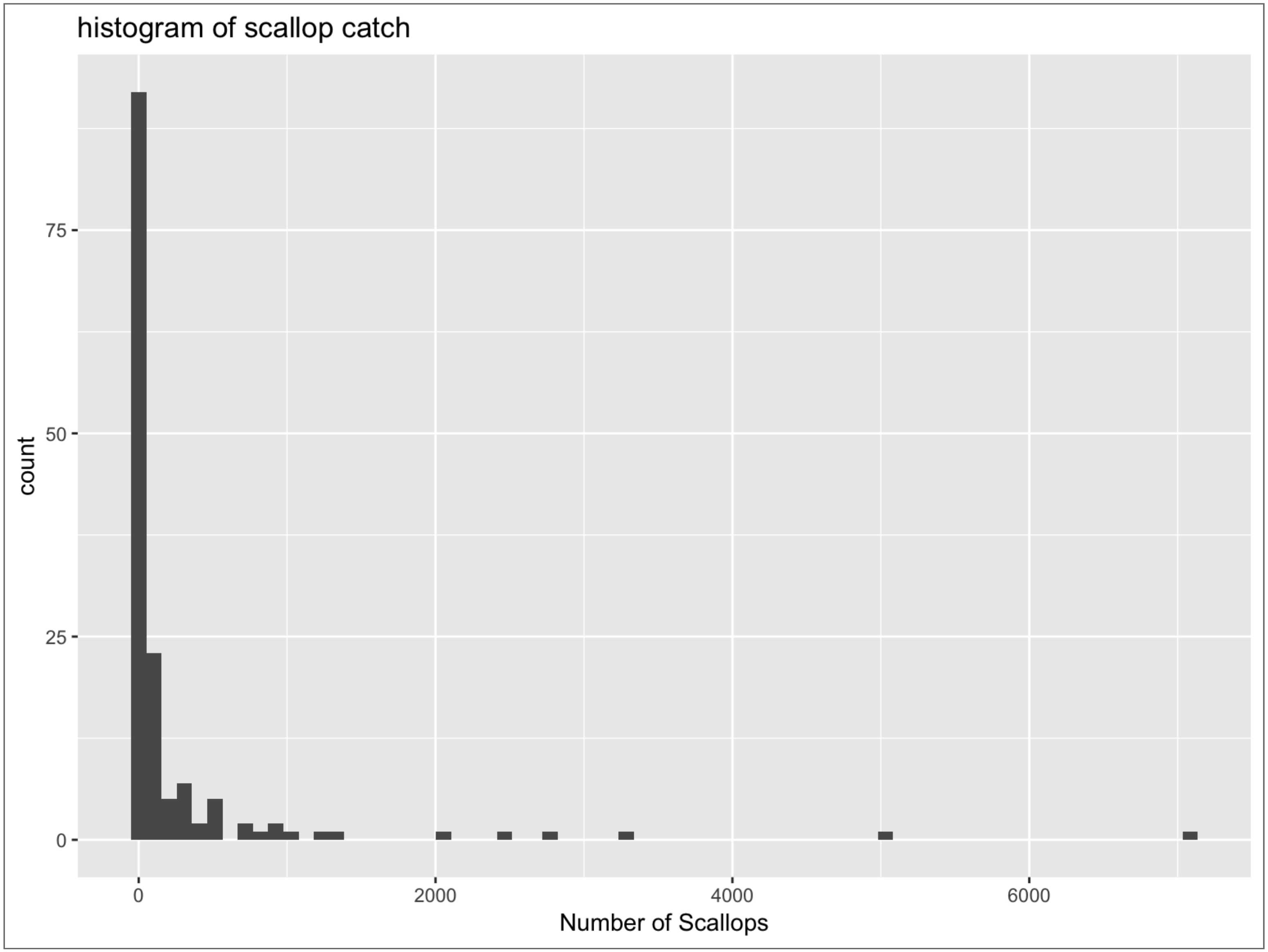
Leaflet

1. Map of Locations - Takeaways

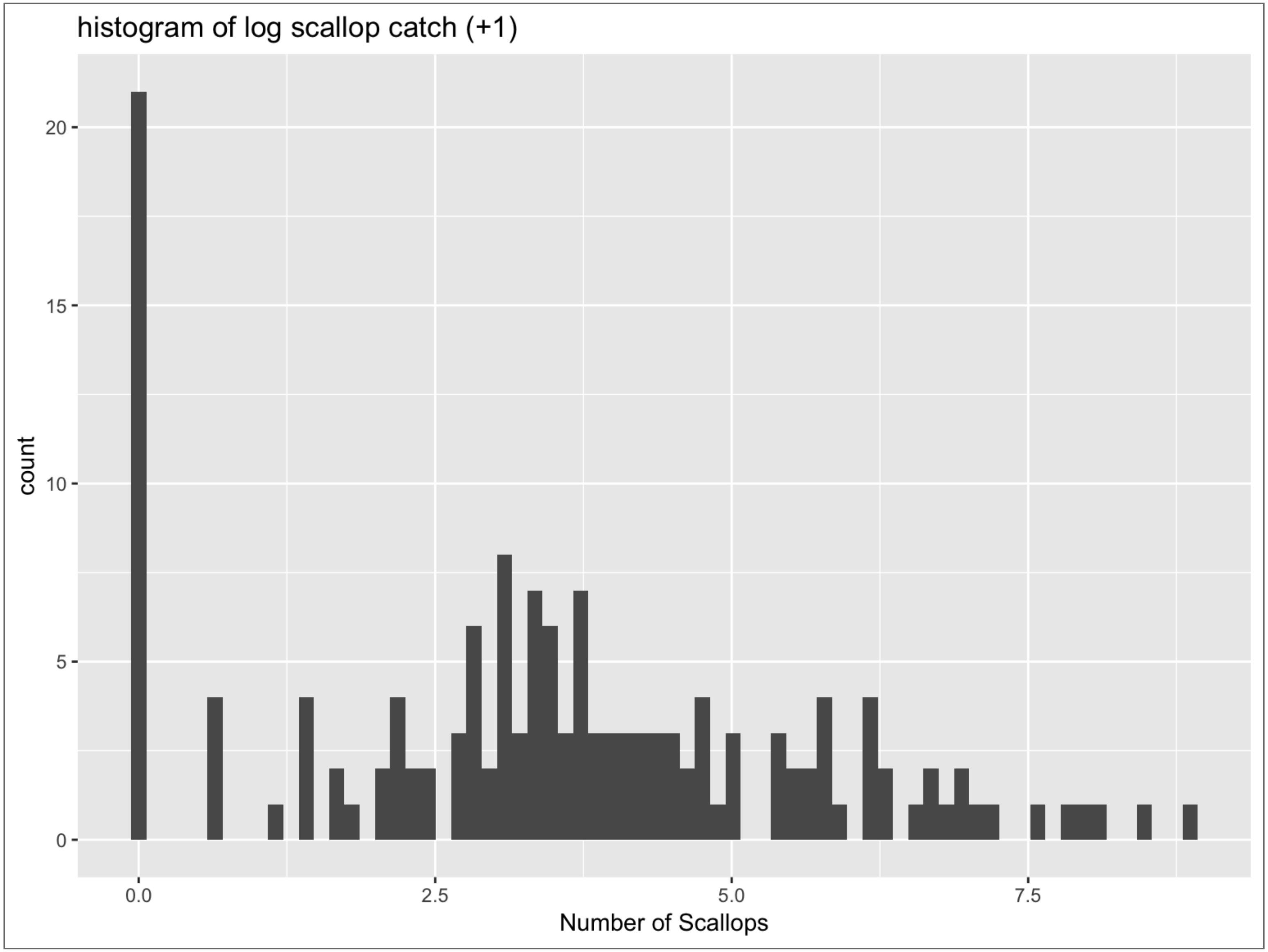
Goal: Understand the sampling approach

- Is this a grid?
- Are there directions that have larger distances?
- How large is the spatial extent?

2. Histogram


histogram of scallop catch
count
75
50
25
0
0
2000
4000
6000
Number of Scallops

2. Histogram

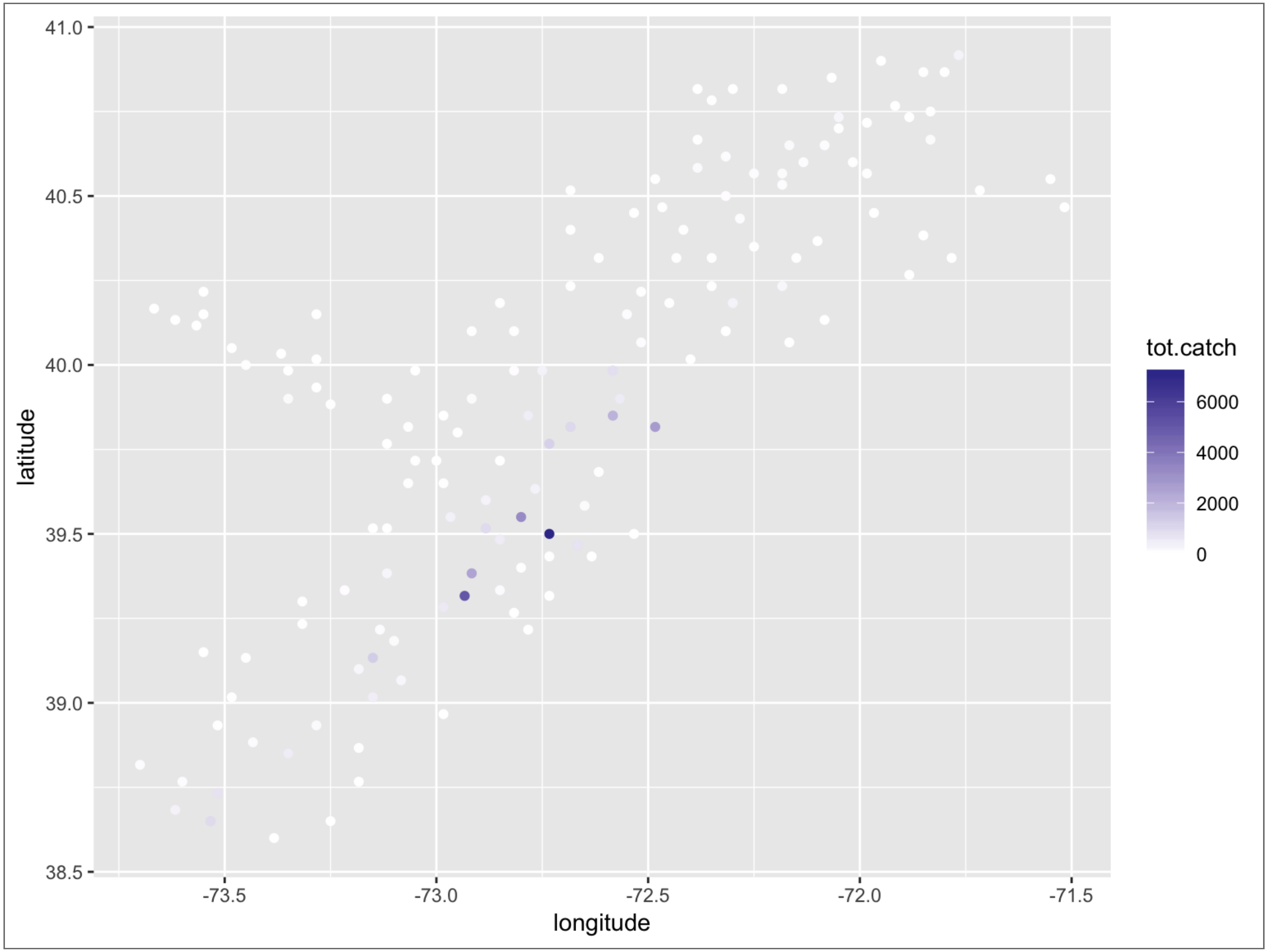

histogram of log scallop catch (+1)
count
20
15
10
5
0
0.0
2.5
5.0
7.5
Number of Scallops

2. Histogram - Takeaways

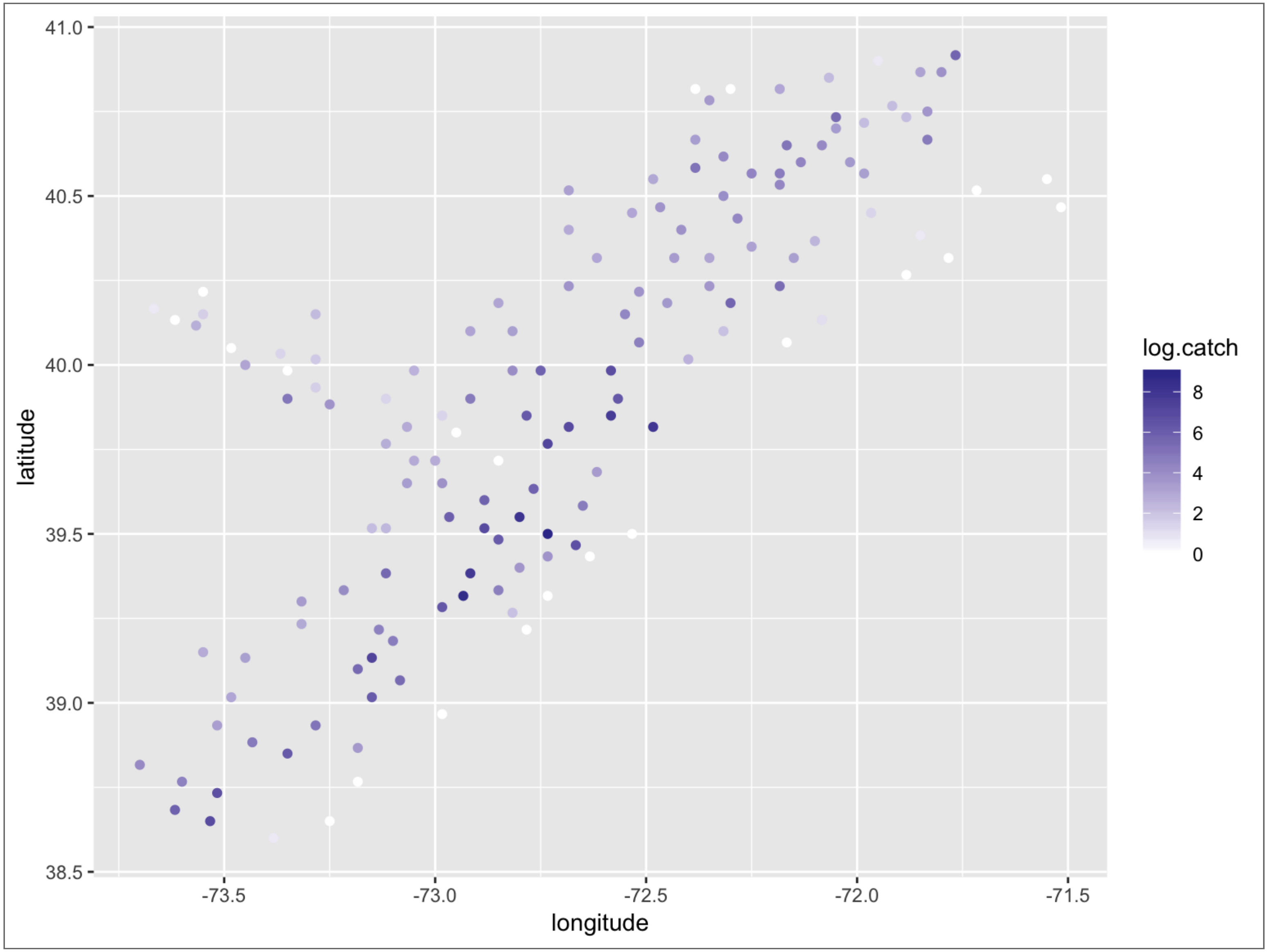
Goal: Identify a sampling distribution for the data

- Continuous or discrete data
- A linear model approach will be used for the response
- Spatial structure can also be included in generalized linear models
- Outliers are worth investigating, but a data point that does not fit the assumed model should not automatically be eliminated

3. 3D scatterplot



3. 3D scatterplot



3. 3D scatterplot - Takeaways

Goal: Examine the spatial pattern of the response

- Again, this is the response not the residual
- Can also think about a contour plot (using some interpolation method)

4. General Regression EDA

- Assessing relationship between variable of interest and covariate information
- No covariates are present in the scallops data

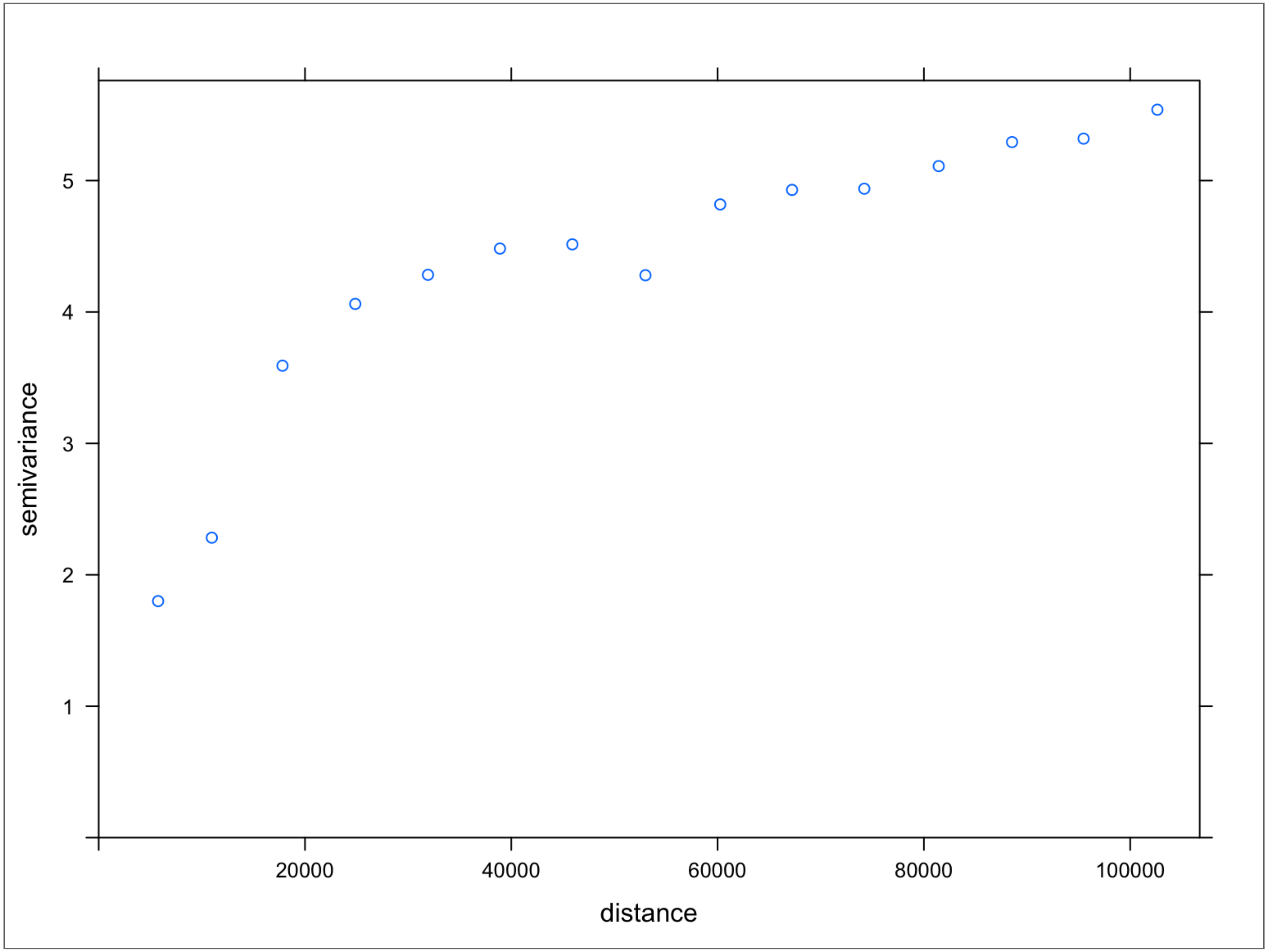
5. Variograms and variogram clouds: Exercise

Explore the code below: what are the differences in the three variograms?

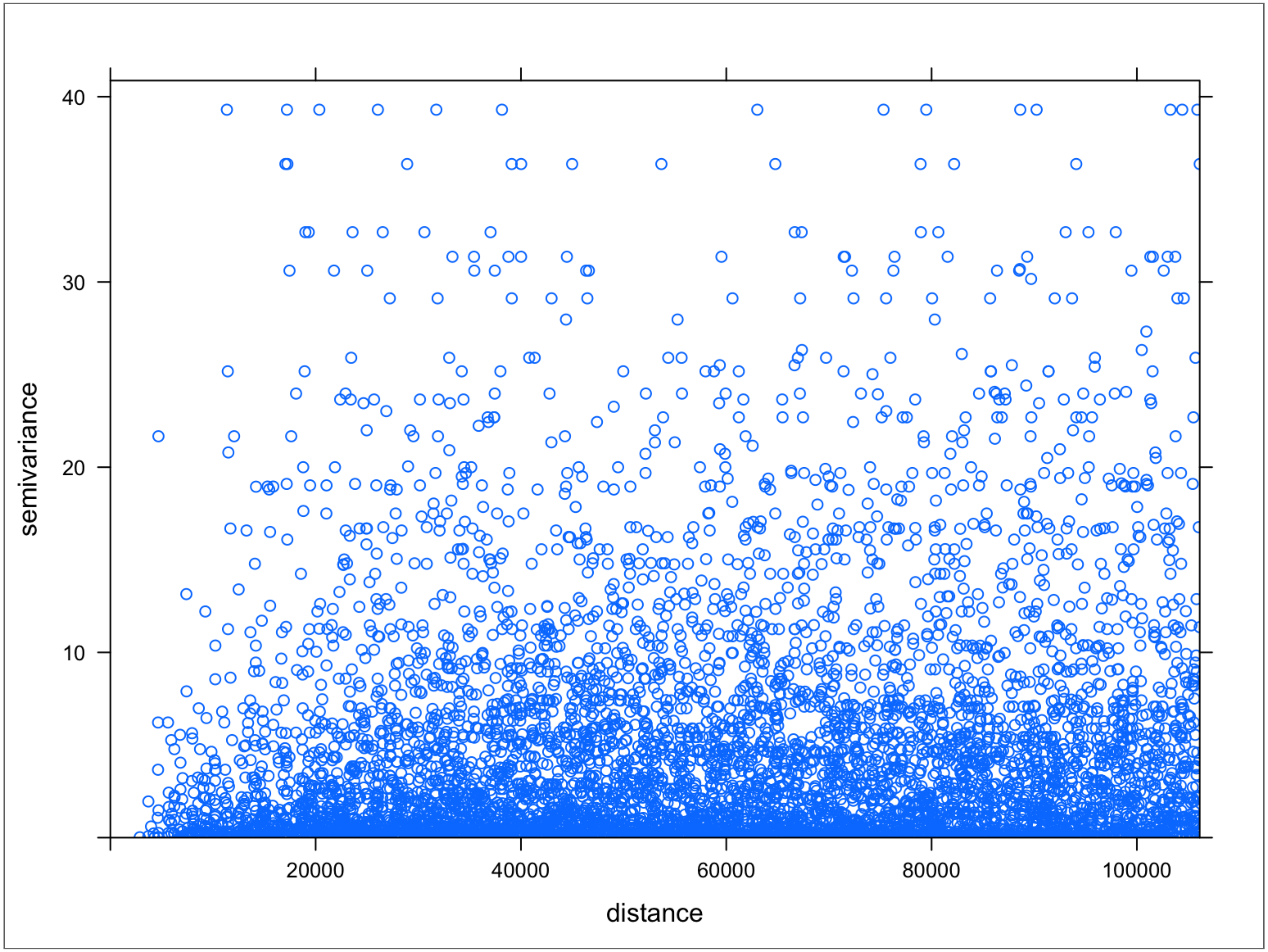
```
coordinates(scallop) = ~longitude+latitude
class(scallop)
scallop.sp <- scallop
proj4string(scallop.sp) <- CRS("+proj=longlat +datum=WGS84")  ## for example
scallop.utm <- spTransform(scallop.sp, CRS("+proj=utm +zone=18 ellps=WGS84"))

plot(variogram(log.catch~1, scallop))
plot(variogram(log.catch~1, scallop.sp))
plot(variogram(log.catch~1, scallop.utm))
```

5. Variograms



5. Variogram Cloud



5. Variograms and variogram clouds: Takeaways

Goal: Visually diagnose spatial structure

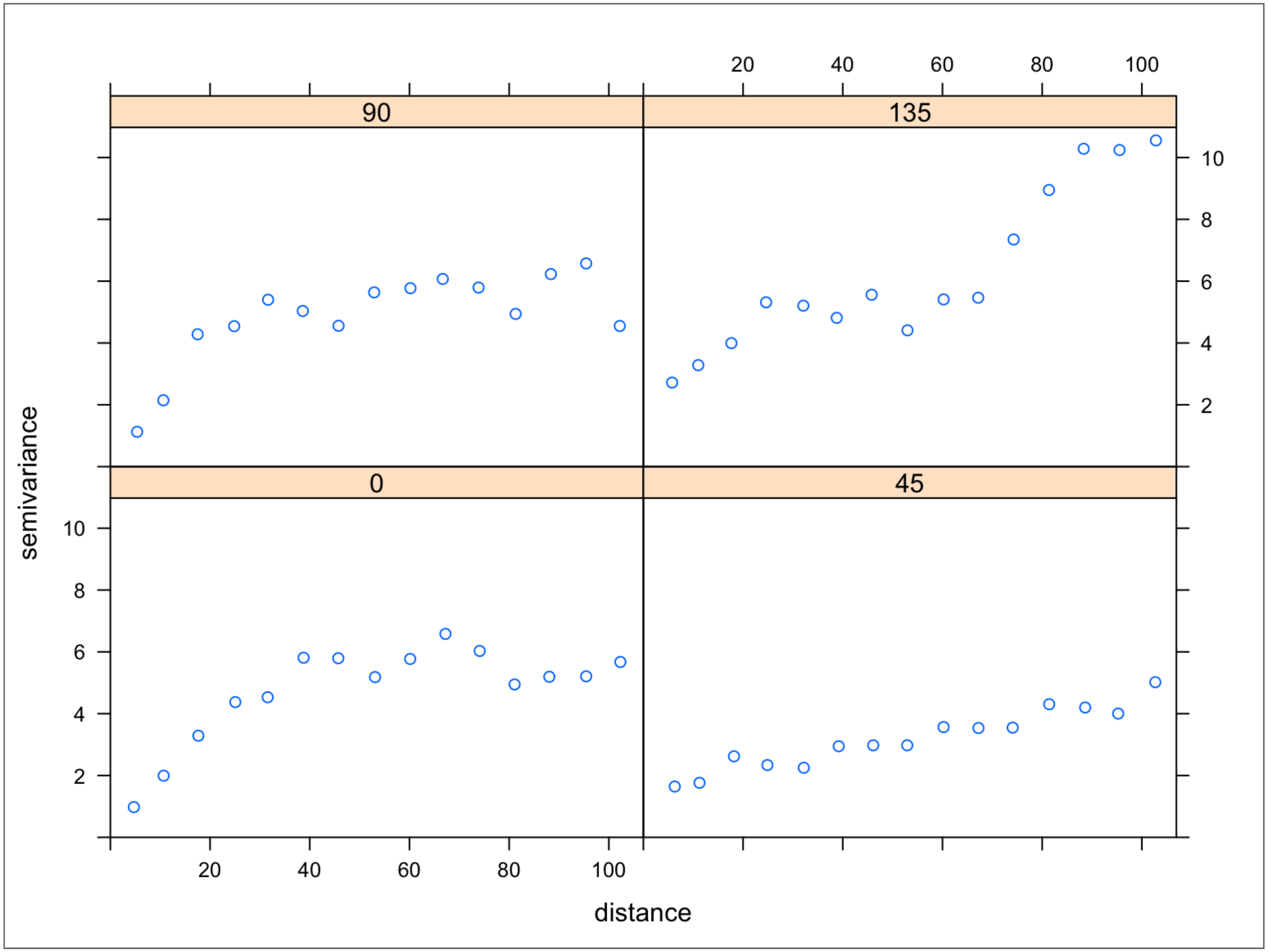
6. Anisotropy

Goal: Determine if direction influencs spatial structure

Anisotropy

Directional Variogram

- All of the variograms we have looked at are isotropic


90
135
0
45
semivariance
distance
20
40
60
80
100
2
4
6
8
10

Separable Correlations Functions

- If the differences in spatial structure are directly related to two coordinate sets, we can create a stationary, anistropic covariance function
- Let
  $cor(Y(\boldsymbol{s}+\boldsymbol{h}), Y(\boldsymbol{s})) = \rho_1(h_y)\rho_2(h_x),$
  where $\rho_1()$ and $\rho_2()$ are proper correlation functions.
- A scaling factor, σ^2, can be used to create covariance.

Geometric Anistropy

- Another solution is the class of geometric anisotropic covariance functions with
  $C(\mathbf{s} - \mathbf{s}') = \sigma^2 \rho((\mathbf{s} - \mathbf{s}')^T B(\mathbf{s} - \mathbf{s}')),$
  where B is positive definite matrix and ρ is a valid correlation function
- B is often referred to as a transformation matrix which rotates and scales the coordinates, such that the resulting transformation can be simplified to a distance.

Sill, Nugget, and Range Anisotropy

- Recall the sill is defined as $\lim_{d \to \infty} \gamma(d)$
- Let $\boldsymbol{h}$ be an arbitrary separation vector, that can be normalized as $\frac{\boldsymbol{h}}{||\boldsymbol{h}||}$
- If $\lim_{a \to \infty} \gamma(a \times \frac{\boldsymbol{h}}{||\boldsymbol{h}||})$ depends on $\boldsymbol{h}$, this is referred to as sill anisotropy.
- Similarly the nugget and range can depend on $\boldsymbol{h}$ and give nugget anisotropy and range anisotropy

Model Fitting

Simulating Spatial Process

- Soon we will look at fitting models for spatial point data
- Simulating data gives a deeper understanding of the model fitting process
- Simulate a mean-zero, isotropic spatial process with a spherical covariance function

Additional Resources

- **Meuse Data Tutorial**
- **Textbook Data Sets**