Spatial EDA

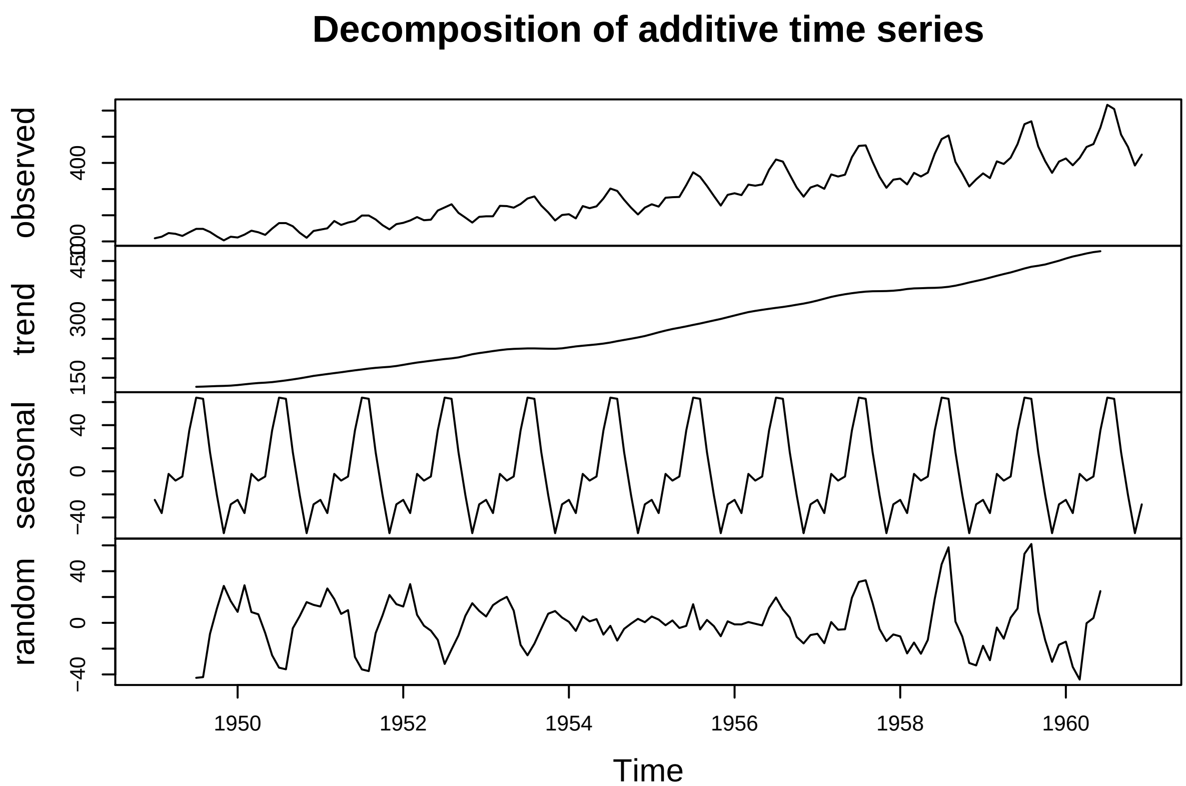
Exploratory Data Analysis

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.



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

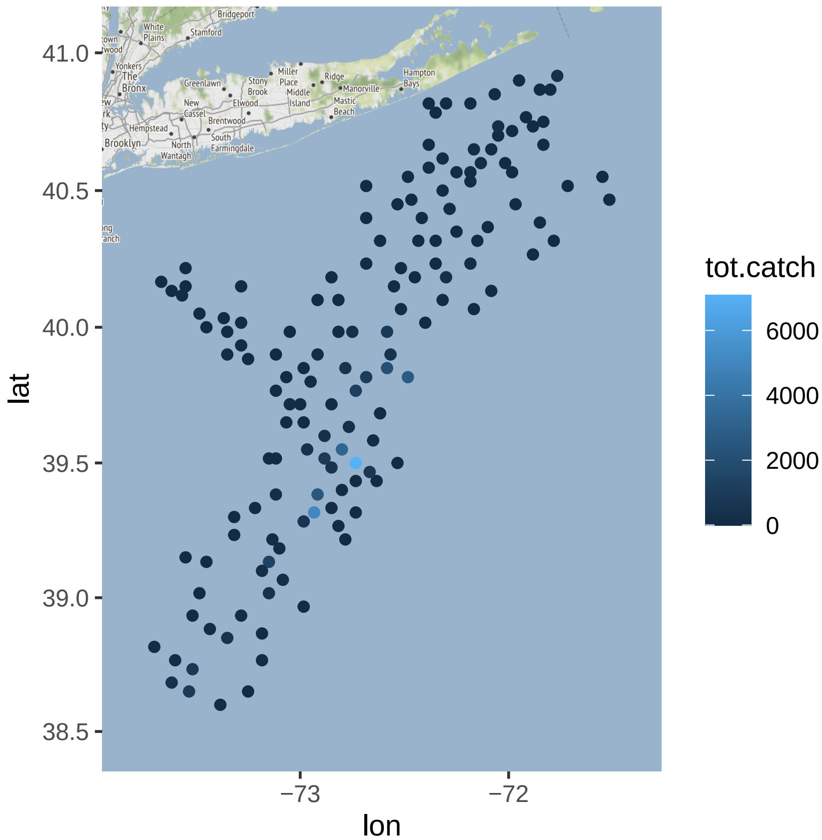
- Spatial structure in the response surface and spatial structure in the error surface are not one-and-the-same.
- $E[(Y(\boldsymbol{s}) - \mu)(Y(\boldsymbol{s'}) - \mu)]$ vs. $E[(Y(\boldsymbol{s}) - \mu(\boldsymbol{s}))(Y(\boldsymbol{s'}) - \mu(\boldsymbol{s}))]$
- There are stationarity implications 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

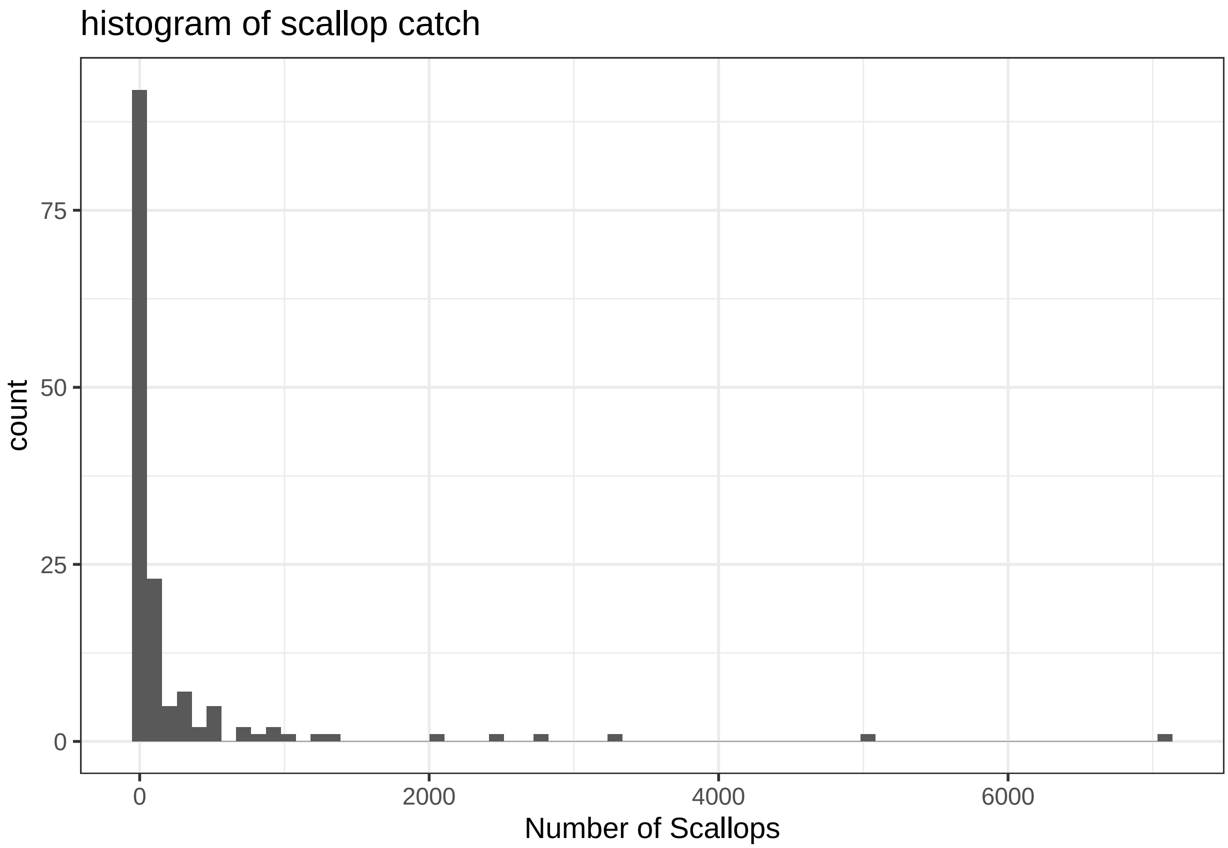


1. Map of Locations - Takeaways

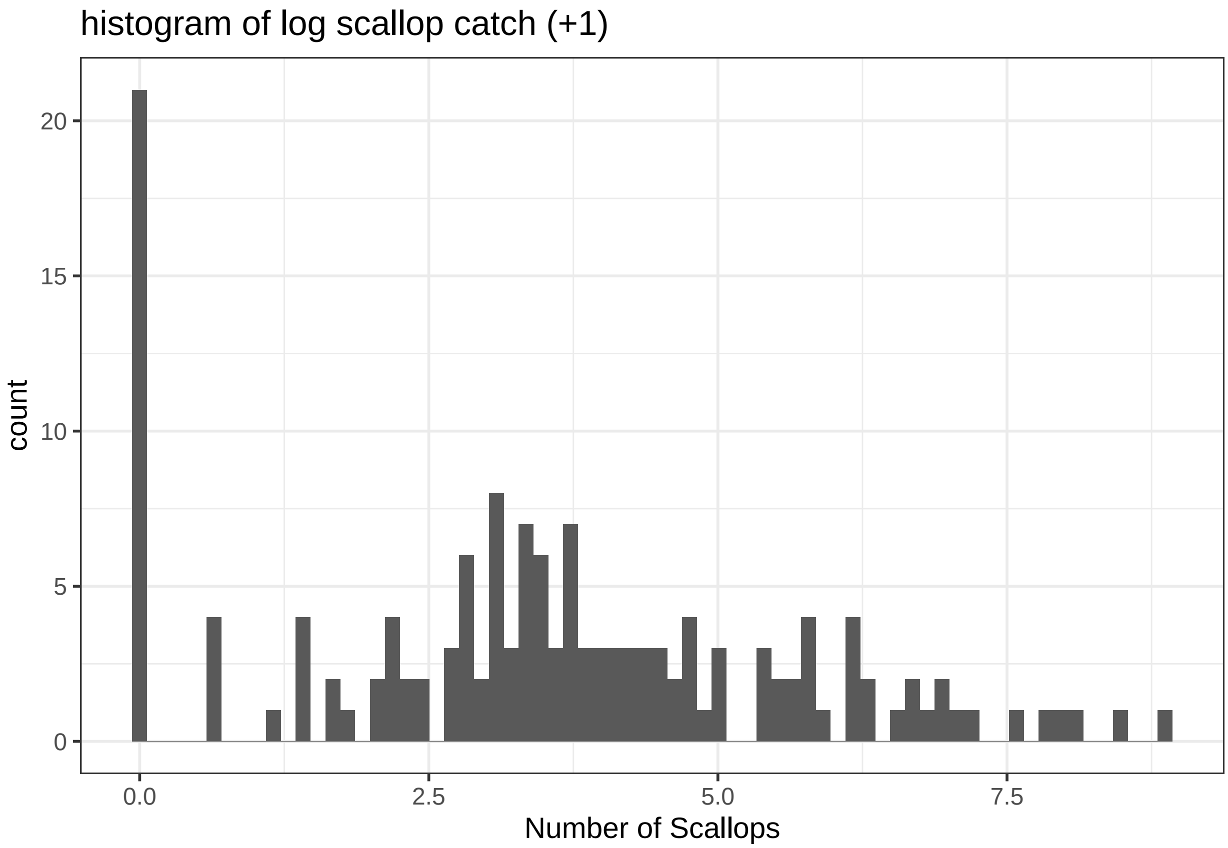
Goal: Understand the sampling approach

- 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

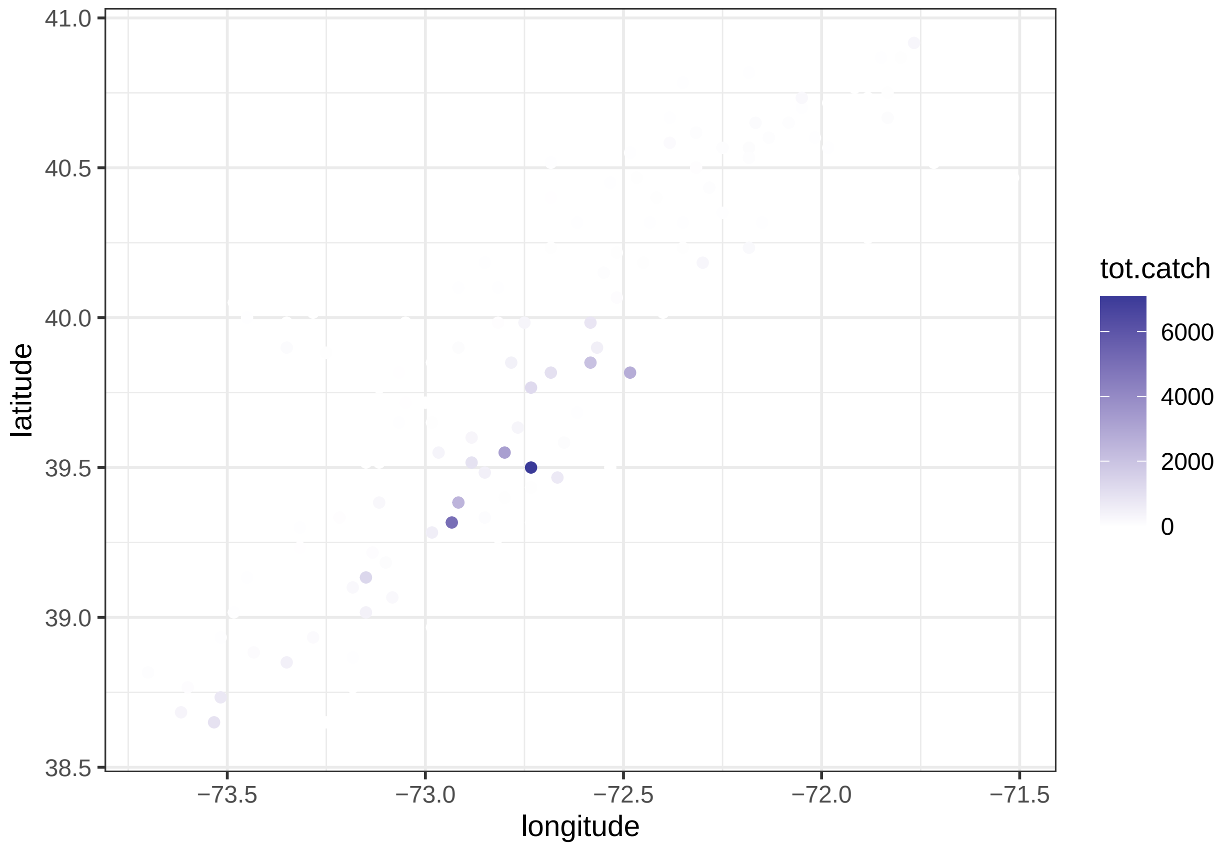


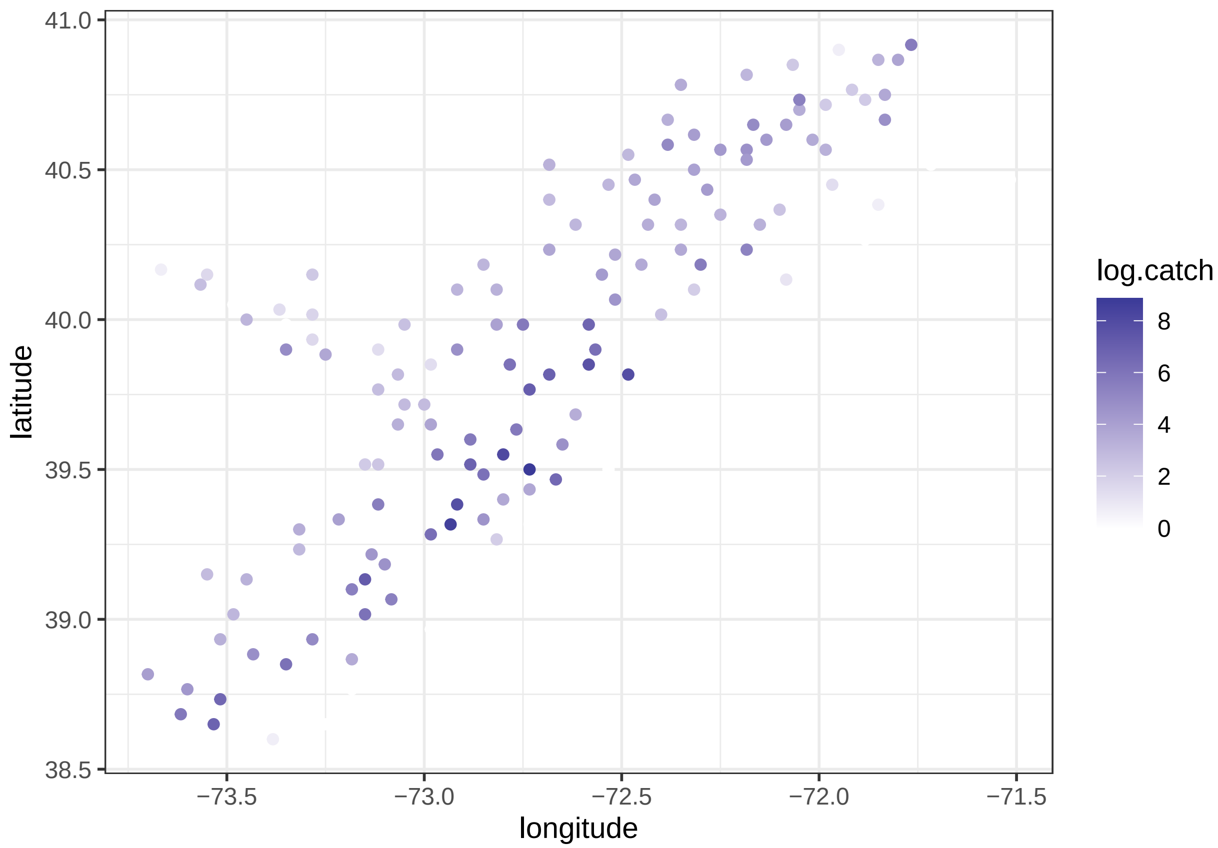
2. Histogram - Takeaways

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 - 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)

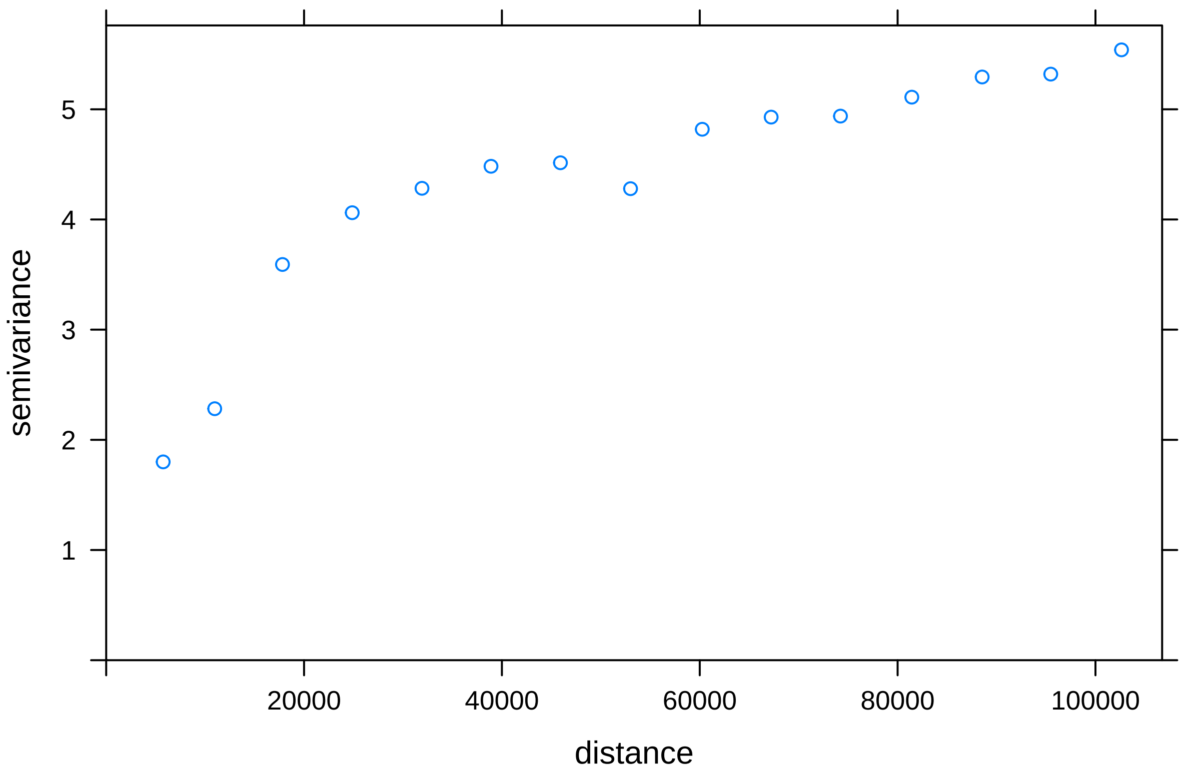
4. General Regression EDA

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

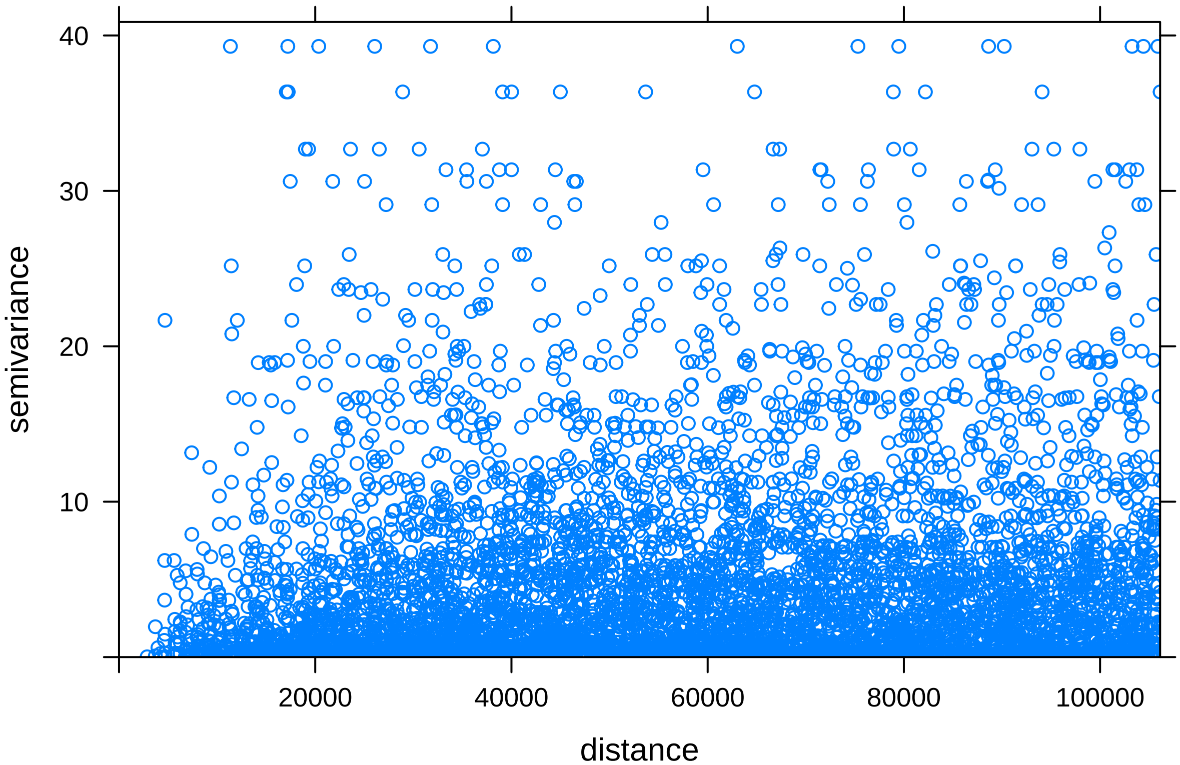
5. Variograms and variogram clouds

5. Variograms

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO", prefer_proj
## = prefer_proj): Discarded datum Unknown based on WGS84 ellipsoid in Proj4
## definition
```



5. Variogram Cloud



5. Variograms and variogram clouds: Takeaways

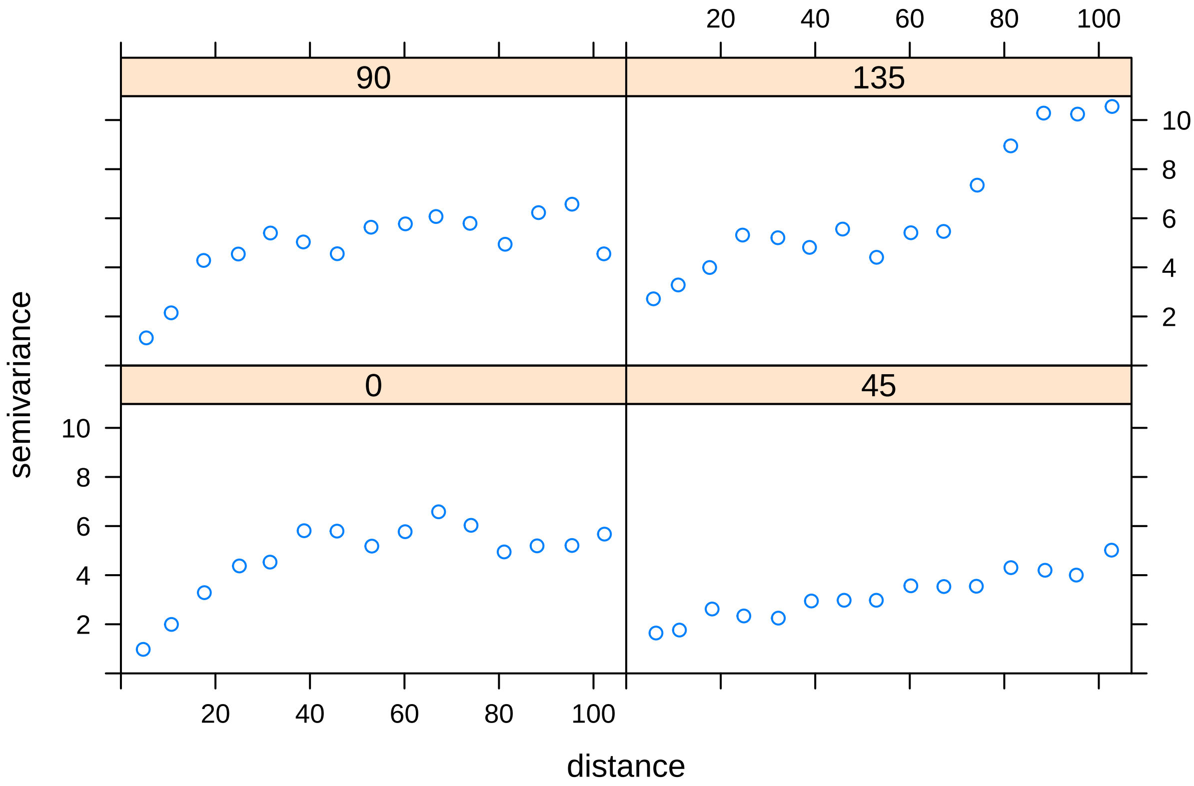
Goal: Visually diagnose spatial structure

6. Anisotropy

Goal: Determine if direction influences spatial structure

Directional Variogram

- All of the variograms we have looked at are isotropic



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
$$C(\boldsymbol{s}-\boldsymbol{s'}) = \sigma^2\rho((\boldsymbol{s}-\boldsymbol{s'})^T B(\boldsymbol{s}-\boldsymbol{s'})),$$
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