Lecture 5: Point Level Models - Variograms

Class Intro

Intro Questions

- What is point referenced data?
- How do we predict the response at unobserved locations?
- For Today:
  - What is a variogram?
  - How is a variogram useful?
  - How is EDA used for point referenced data?

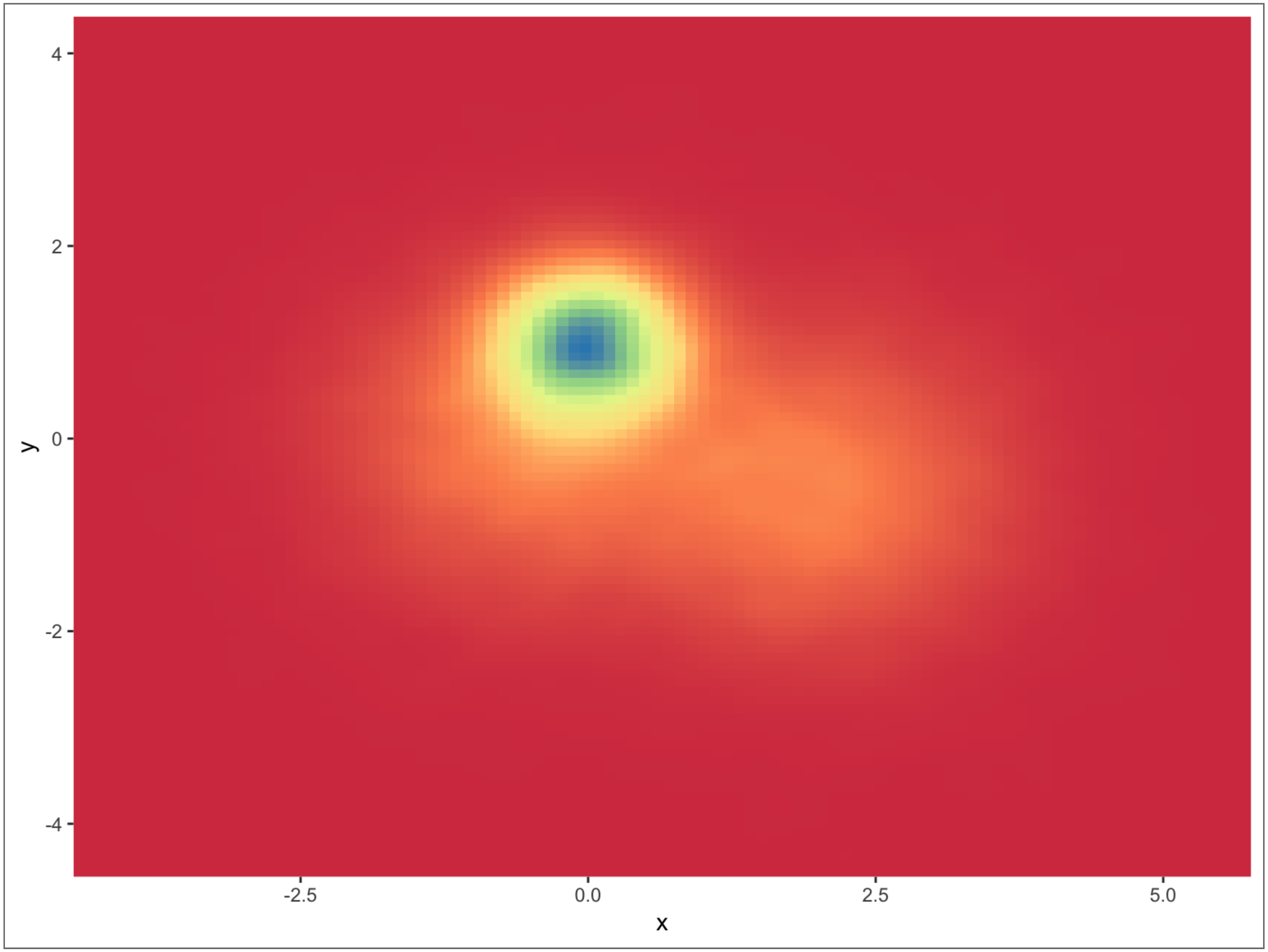
Elements of Geostatistical models: continued…

Stationarity

- Assume the spatial process, $\boldsymbol{Y}(\boldsymbol{s})$ has a mean, $\boldsymbol{\mu}(\boldsymbol{s})$, and that the variance of $\boldsymbol{Y}(\boldsymbol{s})$ exists everywhere.
- The process is **strictly stationary** if: for any $n \geq 1$, any set of n sites $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ and any $\boldsymbol{h} \in \mathcal{R}^r$ (typically r =2), the distribution of $(\boldsymbol{Y}(\boldsymbol{s}_1). \ldots, \boldsymbol{Y}(\boldsymbol{s}+\boldsymbol{h}_n))$ is the same as $\boldsymbol{Y}(\boldsymbol{s}+\boldsymbol{h}_1)$

Strict Stationarity

Is this strictly stationary?



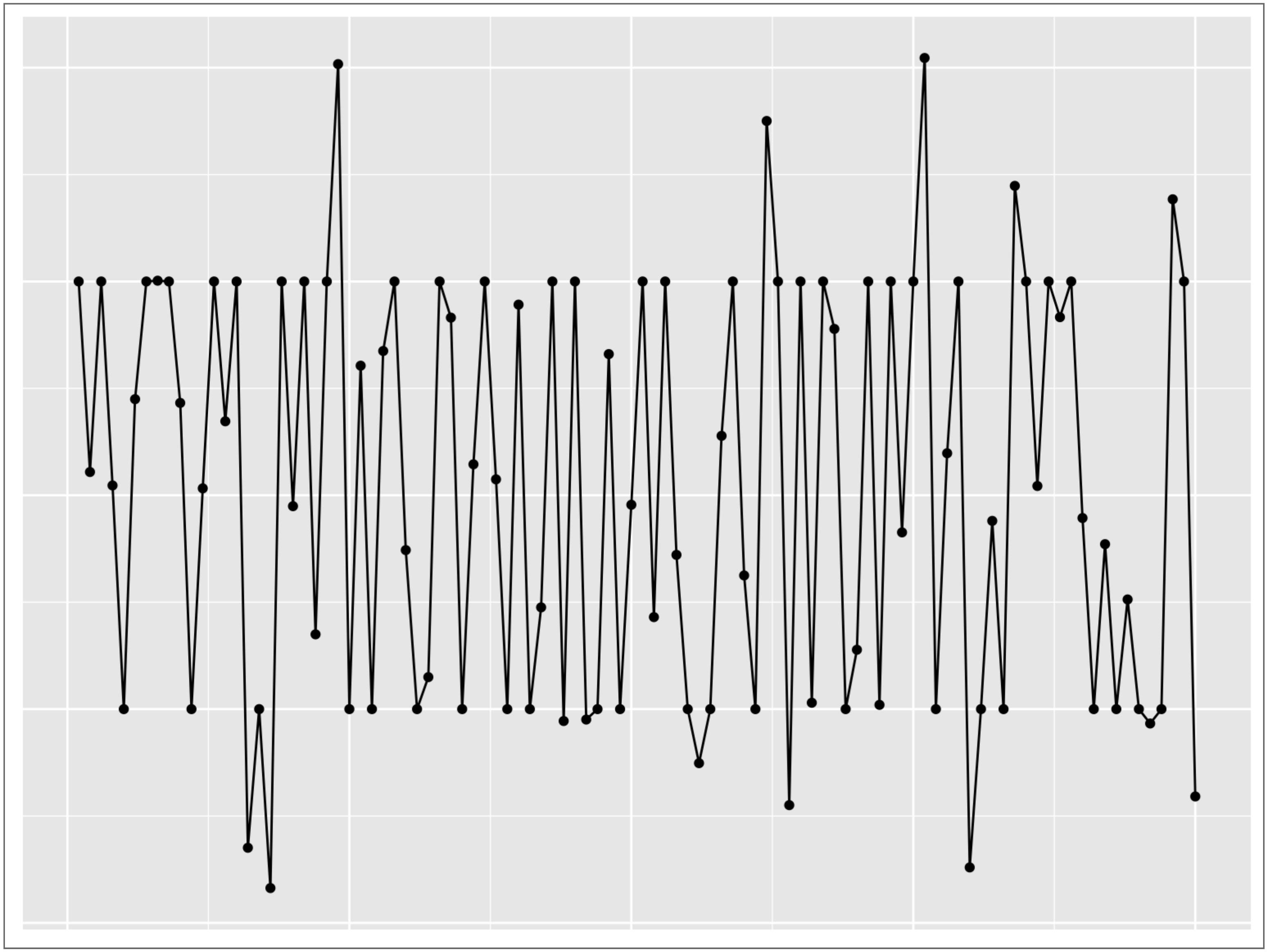
4
2
0
-2
-4
y
-2.5
0.0
2.5
5.0
x

Weak Stationarity

- Weak stationarity, or second-order stationarity of a spatial process, requires a constant mean and covariance that is a function of $\boldsymbol{h}$, $Cov(\boldsymbol{Y}(\boldsymbol{s}), \boldsymbol{Y}(\boldsymbol{s}+\boldsymbol{h})) = C(\boldsymbol{h})$, where $C(\boldsymbol{h})$ is a covariance function that only requires the separation vector $\boldsymbol{h}$.
- Typically with spatial models, the process is assumed to be mean zero as covariates explain the mean structure. So second-order stationarity is primarily focused on the covariance structure.

Weak Stationarity

Is this weakly stationary?



Intrinsic Stationarity

- A third kind of stationarity, known as intrinsic stationarity, describes the behavior of differences in the spatial process, rather than the data.
- Intrinsic stationarity assumes $E[\boldsymbol{Y}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Y}(\boldsymbol{s})] = 0$ then define $E[\boldsymbol{Y}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Y}(\boldsymbol{s})]^2 = Var(\boldsymbol{Y}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Y}(\boldsymbol{s})) = 2\gamma(\boldsymbol{h})$,
  which only works, and satisfies intrinsic stationarity, if the equation only depends on $\boldsymbol{h}$.

Variograms

- Intrinsic stationary justifies the use of variograms $2\gamma(\boldsymbol{h})$
- Variograms are often used to visualize spatial patterns:
  - If the distance between points is small, the variogram is expected to be small
  - As the distance between points increases, the variogram increases
- There is a mathematical link between the covariance function $C(\boldsymbol{h})$ and the variogram $2\gamma(\boldsymbol{h})$.

Isotropy

- If the variogram $2\gamma(\boldsymbol{h})$ depends only on the length of $\boldsymbol{h}$, $||\boldsymbol{h}||$, and not the direction, then the variogram is isotropic.
- If the direction of $\boldsymbol{h}$ impacts the variogram, then the variogram is anisotropic.

More about Variograms

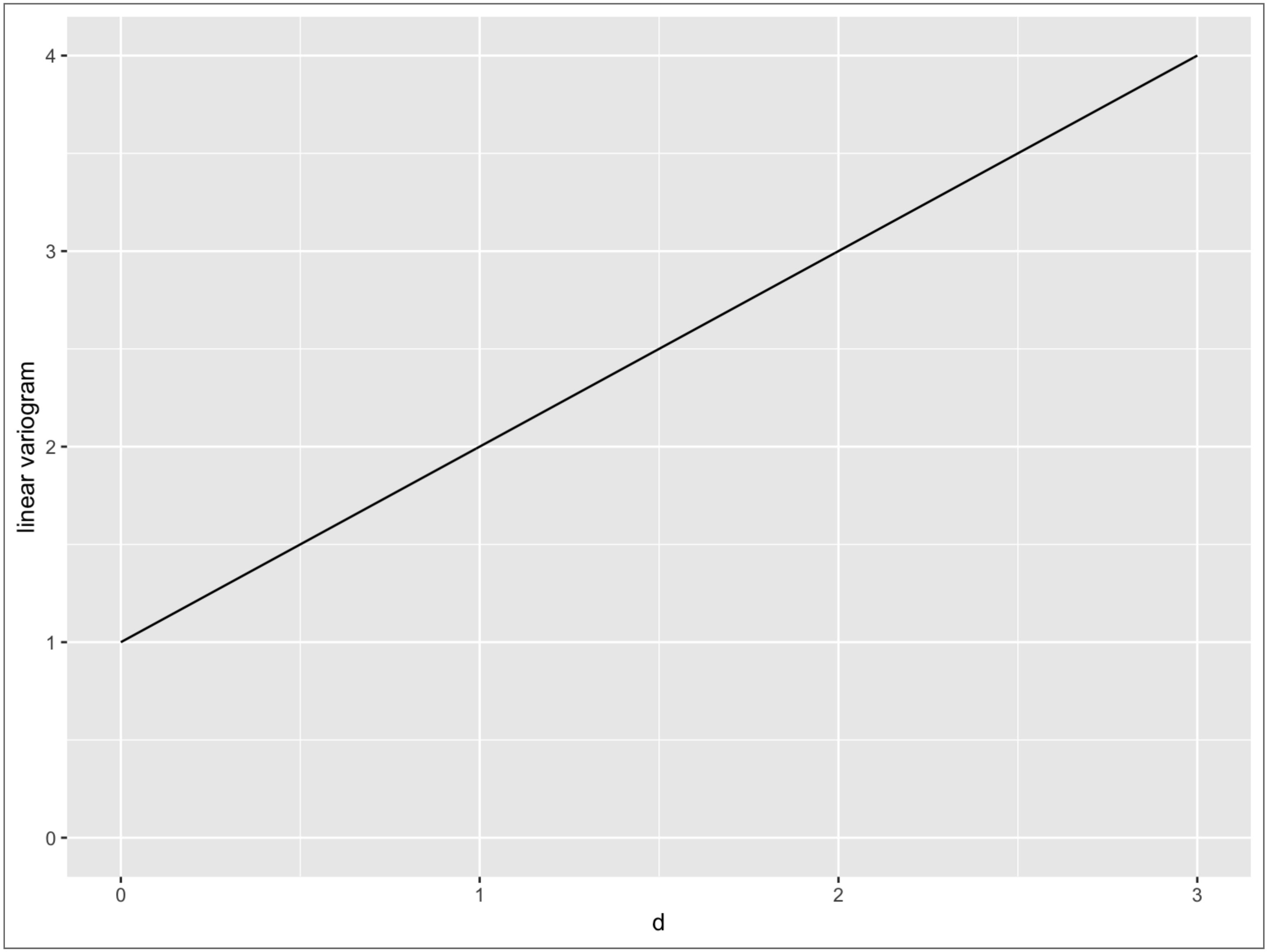
Variogram to the Covariance Function

$$
\begin{aligned}
2\gamma(\boldsymbol{h}) &= Var(\boldsymbol{Y}(\mathbf{s}+\boldsymbol{h}) - \boldsymbol{Y}(\mathbf{s})) \\
&= Var(\boldsymbol{Y}(\mathbf{s}+\boldsymbol{h})) + Var(\boldsymbol{Y}(\mathbf{s})) - 2Cov(Var(\boldsymbol{Y}(\mathbf{s}+\boldsymbol{h}), \boldsymbol{Y}(\mathbf{s}))) \\
&= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\boldsymbol{h}) \\
&= 2\left[C(\mathbf{0}) - C(\boldsymbol{h})\right]
\end{aligned}
$$

- Given $C()$, the variogram can easily be recovered
- Going the other way, $2\gamma() \rightarrow C()$, requires additional assumptions

Linear semivariogram

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 d & \text{if } d > 0 \\ 0 & \text{otherwise} \end{cases}$$


linear variogram
0
1
2
3
4
0
1
2
3
d

Nugget, sill, partial-sill, and range

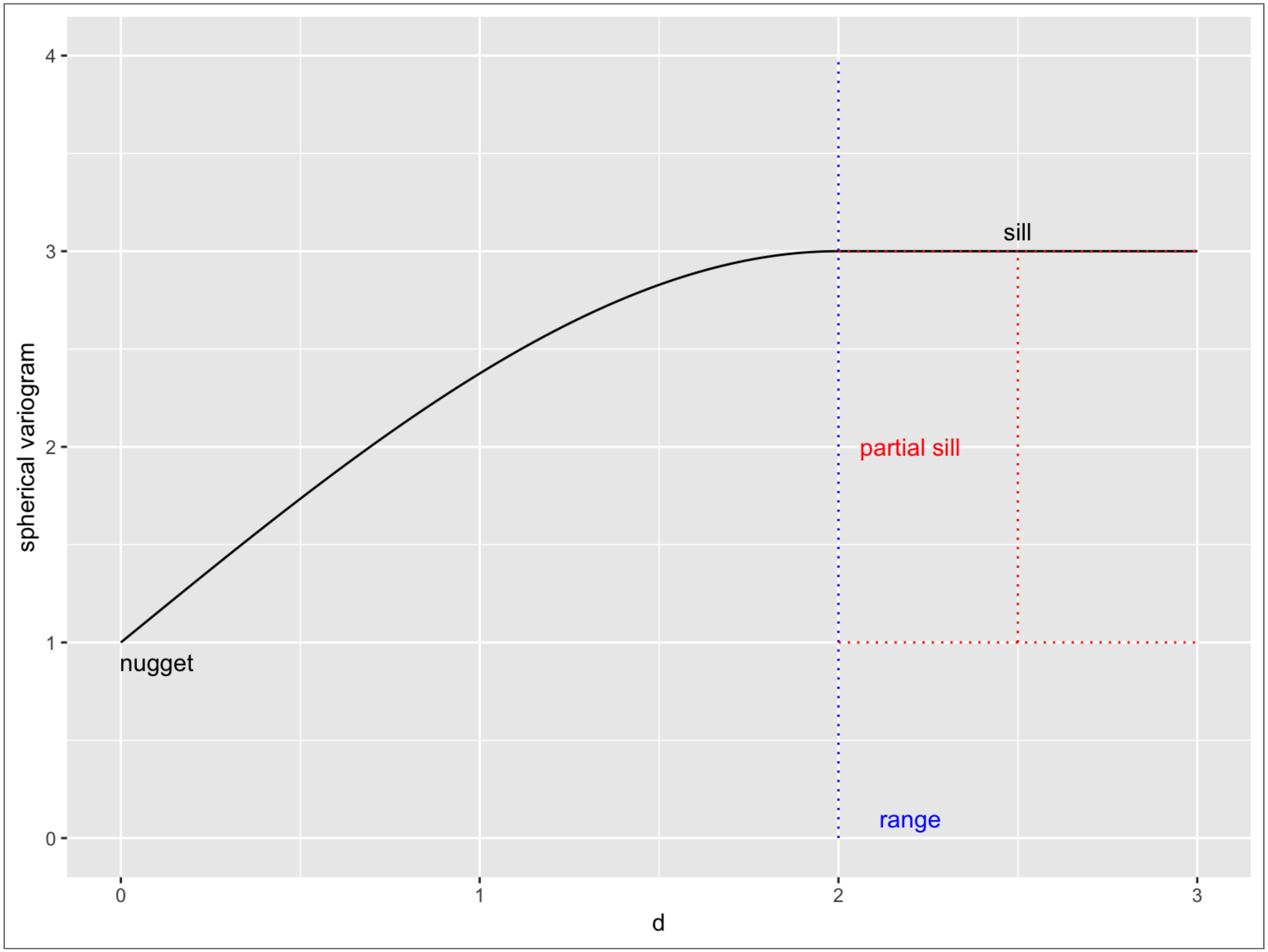
- Nugget is defined as $\gamma(d): \; d \to 0_+$
- Sill is defined as $\gamma(d): \; d \to \infty$
- Partial sill is defined as the $\text{sill} - \text{nugget}$
- Range is defined as the first point where $\gamma(d)$ reaches the sill. Note the Range is often defined as $R = \frac{1}{\phi}$, but confusingly both R and ϕ are sometimes referred to as the range. More properly, ϕ is called the decay parameter.

Spherical semivariogram: Exercise

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 & \text{if } d \geq 1/\phi \\ \tau^2 + \sigma^2 \left[\frac{3\phi d}{2} - \frac{1}{2}(\phi d)^3 \right] \\ 0 & \text{otherwise} \end{cases}$$

- Sketch, or generate in R, a spherical semivariogram
- On this figure label the nugget, sill, partial sill, and range.

Spherical semivariogram: Solution



More Variograms

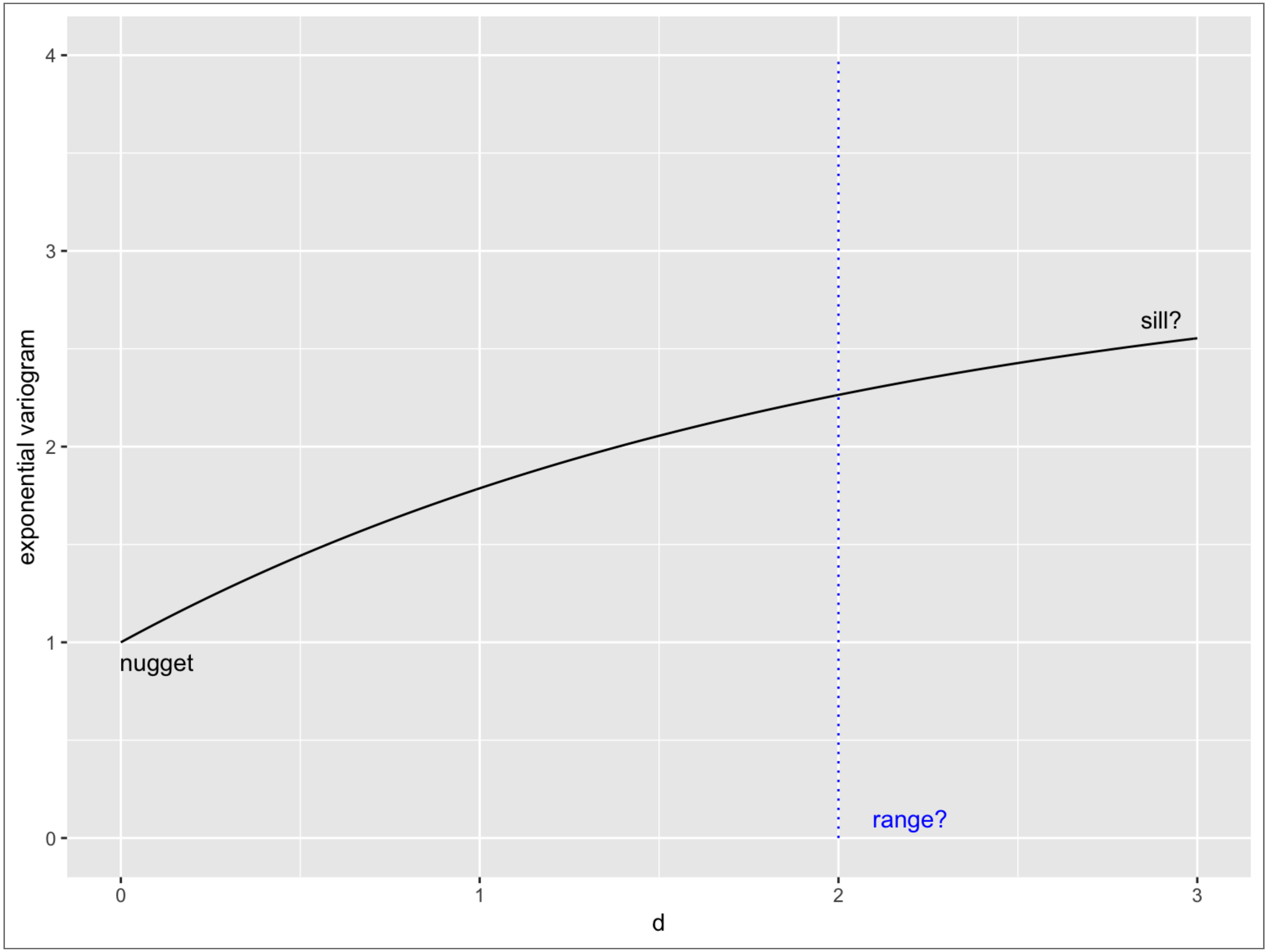
Exponential

- We saw the exponential covariance earlier in class, what is the mathematical form?
- $C(d) = \begin{cases} \tau^2 + \sigma^2 & \text{if } d = 0 \\ \sigma^2 \exp(-\phi d) & \text{if } d > 0 \end{cases}$
- The variogram is
  $\gamma(d) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi d)) & \text{if } d > 0 \\ 0 & \text{otherwise} \end{cases}$

Exponential Semivariogram: Exercise

- Plot the exponential semivariogram
- Identify and interpret: the nugget, range, and sill. How are they similar or different from the spherical semivariogram?

Exponential Semivariogram: Solution

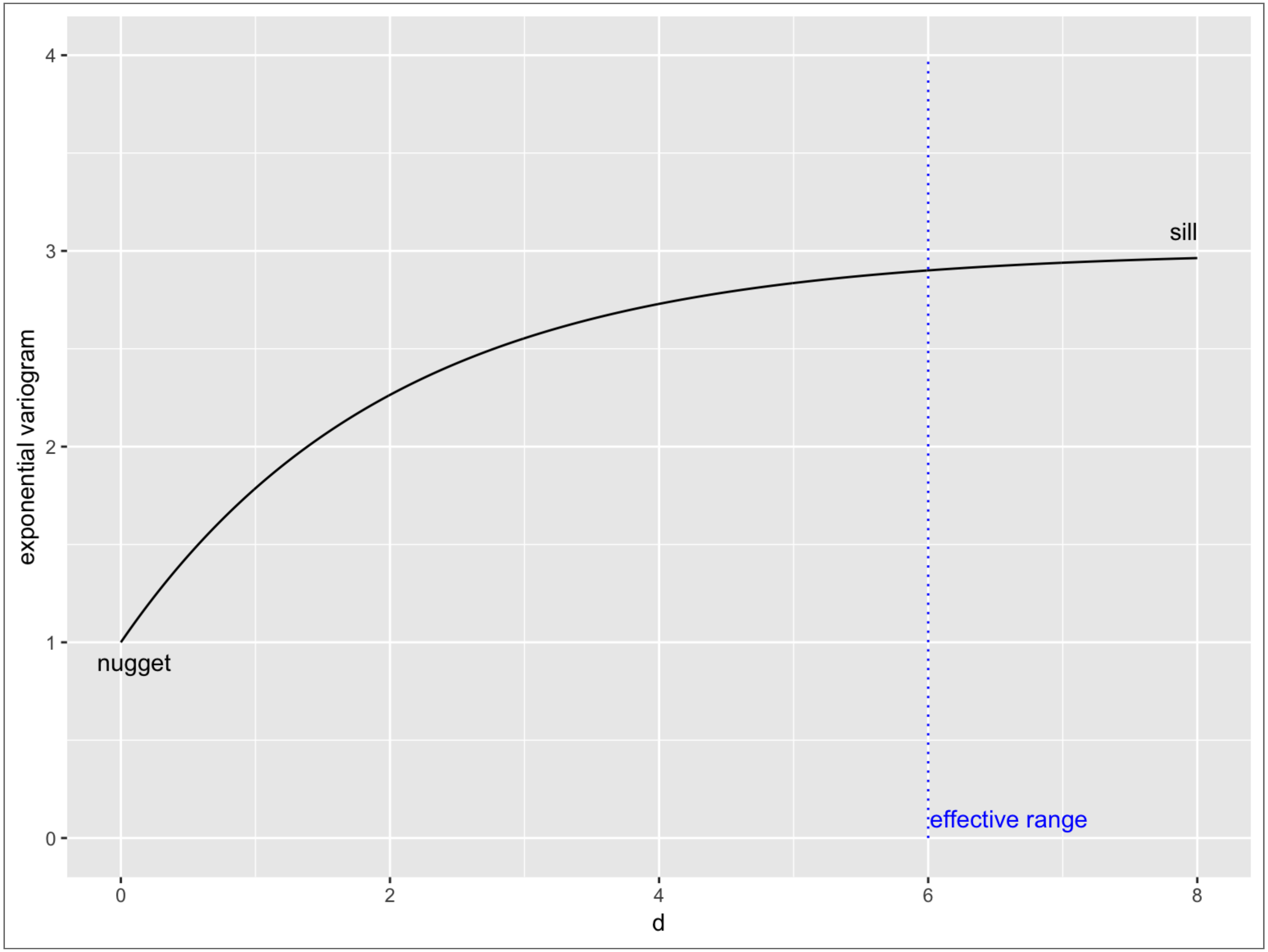


Exponential Semivariogram: Solution 2

- The sill is only reached asymptotically when $\lim_{d \to \infty} exp(-\phi * d) \to 0$
- Thus, the range, defined as distance where semivariogram reaches the sill is ∞
- The *effective range* is defined as the distance where there is *effectively* no spatial structure. Generally this is determined when the correlation is .05.
- $\exp(-d_o \phi) = .05$

$$d_0 \phi = -\log(.05)$$
$$d_0 \approx \frac{3}{\phi}$$

Exponential Semivariogram: Solution 3



More Semivariograms: Equations

- Gaussian:
$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi^2 d^2)) & \text{if } d > 0 \\ 0 & \text{otherwise} \end{cases}$$
- Powered Exponential:
$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-|\phi d|^p)) & \text{if } d > 0 \\ 0 & \text{otherwise} \end{cases}$$
- Matérn:
$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 \left[1 - \frac{(2\sqrt{\nu} d\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(2\sqrt{\nu} d\phi)\right] & \text{if } d > 0 \\ 0 & \text{otherwise} \end{cases},$$
where K_ν is a modified Bessel function and $\Gamma()$ is a Gamma function.

Variogram Creation

Variogram Creation: How?

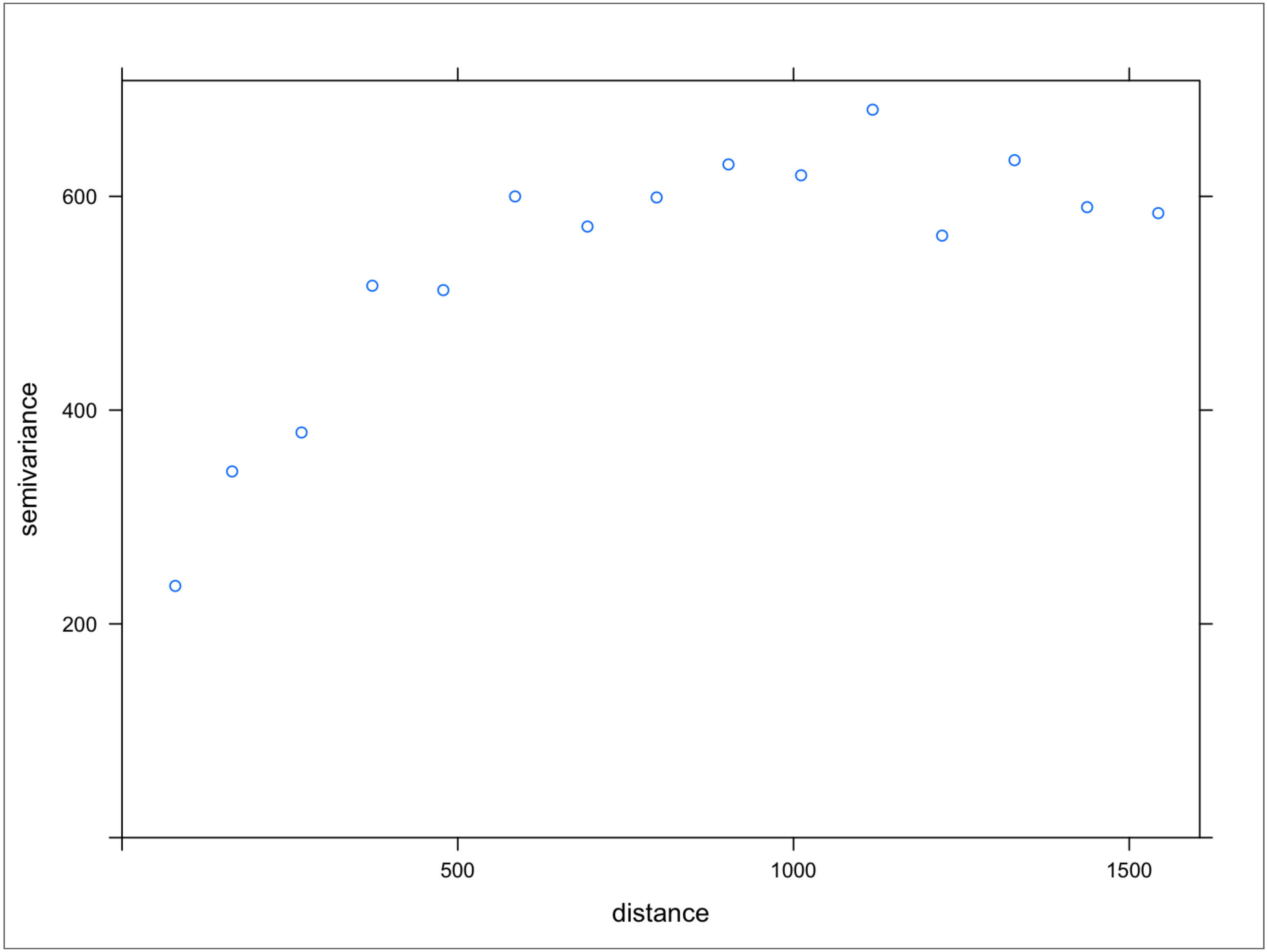
x	y	copper
181072	333611	85
181025	333558	81
181165	333537	68
181298	333484	81
181307	333330	48
181390	333260	61
181165	333370	31
181027	333363	29
181060	333231	37
181232	333168	24
181191	333115	25
181032	333031	25
180874	333339	93

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
```

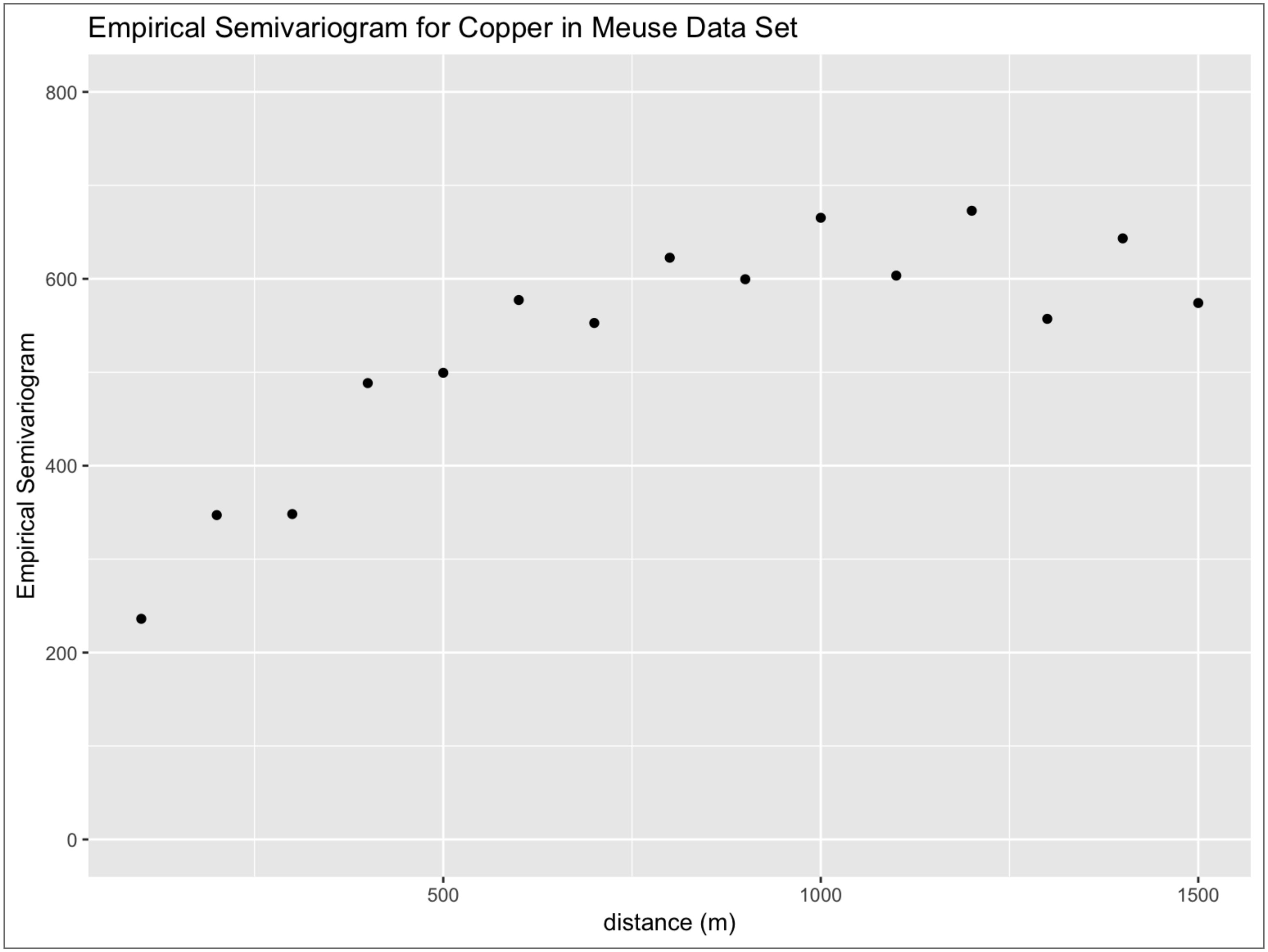
Variogram Creation: How - Step 3

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
0
200
400
600
800
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??