Stat 534 Homework 5

Kenny Flagg

February 17, 2017

1. *Let $\gamma(\mathbf{s}_i, s_j) = \gamma(\mathbf{h}_{ij})$ be a semivariogram for a second-order stationary spatial process.*

   (a) *Show that $\gamma(\mathbf{h}_{ij}) = C(\mathbf{0}) - C(\mathbf{h}_{ij})$.*

$$\begin{aligned}\gamma(\mathbf{h}_{ij}) &= \frac{1}{2}Var\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right]\\ &= \frac{1}{2}\left[Var(Z(\mathbf{s}_i)) + Var(Z(\mathbf{s}_j)) - 2Cov(Z(\mathbf{s}_i), Z(\mathbf{s}_j))\right]\\ &= \frac{1}{2}Var(Z(\mathbf{s}_i)) + \frac{1}{2}Var(Z(\mathbf{s}_j)) - Cov(Z(\mathbf{s}_i), Z(\mathbf{s}_j))\\ &= \frac{1}{2}C(\mathbf{0}) + \frac{1}{2}C(\mathbf{0}) - C(\mathbf{s}_i - \mathbf{s}_j)\\ &= C(\mathbf{0}) - C(\mathbf{h}_{ij})\end{aligned}$$

   (b) *Show that $\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(\mathbf{h}_{ij}) \leq 0$ for any sites $\mathbf{s}_i$, $i = 1, \dots, n$ and for constants a_i, $i = 1, \dots, n$ with $\sum_{i=1}^{n} a_i = 0$.*

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(\mathbf{h}_{ij}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j [C(\mathbf{0}) - C(\mathbf{h}_{ij})]\\ &= C(\mathbf{0})\sum_{i=1}^{n} a_i \sum_{j=1}^{n} a_j - \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(Z(\mathbf{s}_i), Z(\mathbf{s}_j))\\ &= 0 - \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left[E(Z(\mathbf{s}_i)Z(\mathbf{s}_j)) - E(Z(\mathbf{s}_i))E(Z(\mathbf{s}_j))\right]\\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j E(Z(\mathbf{s}_i)Z(\mathbf{s}_j)) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j E(Z(\mathbf{s}_i))E(Z(\mathbf{s}_j))\\ &= -E\left(\sum_{i=1}^{n} a_i Z(\mathbf{s}_i) \sum_{j=1}^{n} a_j Z(\mathbf{s}_j)\right) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mu^2\\ &= -E\left(\sum_{i=1}^{n} a_i Z(\mathbf{s}_i)\right)^2 + 0\\ &\leq 0\end{aligned}$$

2. *Matheron's semivariogram estimator is*

$$\widehat{\gamma}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\}^2$$

where $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{h} = \mathbf{s}_i - \mathbf{s}_j\}$ *and* $|N(\mathbf{h})|$ *is the number of pairs in the set* $N(\mathbf{h})$*.*

(a) *Let* $Z(\mathbf{s}) = \mu + e(\mathbf{s})$ *where* $E[e(\mathbf{s})] = 0$ *with* $\gamma_Z(\mathbf{h}) = \gamma_e(\mathbf{h})$ *(adding a constant to the random error terms does not change the variance/covariance properties of the process). Show that* $\widehat{\gamma}(\mathbf{h})$ *is unbiased for* $\gamma_Z(\mathbf{h})$*. That is, show* $E[\widehat{\gamma}(\mathbf{h})] = \gamma_Z(\mathbf{h})$*. Hint: Show that under an assumption of a constant mean*

$$\gamma_Z(\mathbf{h}) = \frac{1}{2} E\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right]^2$$

First,

$$\begin{aligned}
\gamma_Z(\mathbf{h}) &= \frac{1}{2} Var\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right] \\
&= \frac{1}{2}\left(E\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right]^2 - \left(E[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]\right)^2\right) \\
&= \frac{1}{2}\left(E\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right]^2 - \left(E[Z(\mathbf{s}_i)] - E[Z(\mathbf{s}_j)]\right)^2\right) \\
&= \frac{1}{2}\left(E\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right]^2 - (\mu - \mu)^2\right) \\
&= \frac{1}{2} E\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right]^2 .
\end{aligned}$$

Then,

$$\begin{aligned}
E\left[\widehat{\gamma}(\mathbf{h})\right] &= E\left[\frac{1}{2|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\}^2\right] \\
&= \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \frac{1}{2} E\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right]^2 \\
&= \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \gamma_Z(\mathbf{h}_{ij}) \\
&= \frac{1}{|N(\mathbf{h})|} |N(\mathbf{h})| \gamma_Z(\mathbf{h}) \\
&= \gamma_Z(\mathbf{h})
\end{aligned}$$

so $\widehat{\gamma}(\mathbf{h})$ is unbiased for $\gamma_Z(\mathbf{h})$.

(b) *We pointed out in class that Matheron's Estimator is biased in the presence of trend. Let $Z(\mathbf{s}) = \mu(\mathbf{s}) + e(\mathbf{s})$ with $E[e(\mathbf{s})] = 0$ and $\gamma_Z(\mathbf{h}) = \gamma_e(\mathbf{h})$. Show*

$$E\left[\widehat{\gamma}(\mathbf{h})\right] = \gamma_e(\mathbf{h}) + \frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}\left[\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)\right]^2.$$

First note that

$$\gamma_e(\mathbf{h}) = \frac{1}{2}E\left[e(\mathbf{s}_i) - e(\mathbf{s}_j)\right]^2$$

because the mean of $e(\mathbf{s})$ is constant. Now,

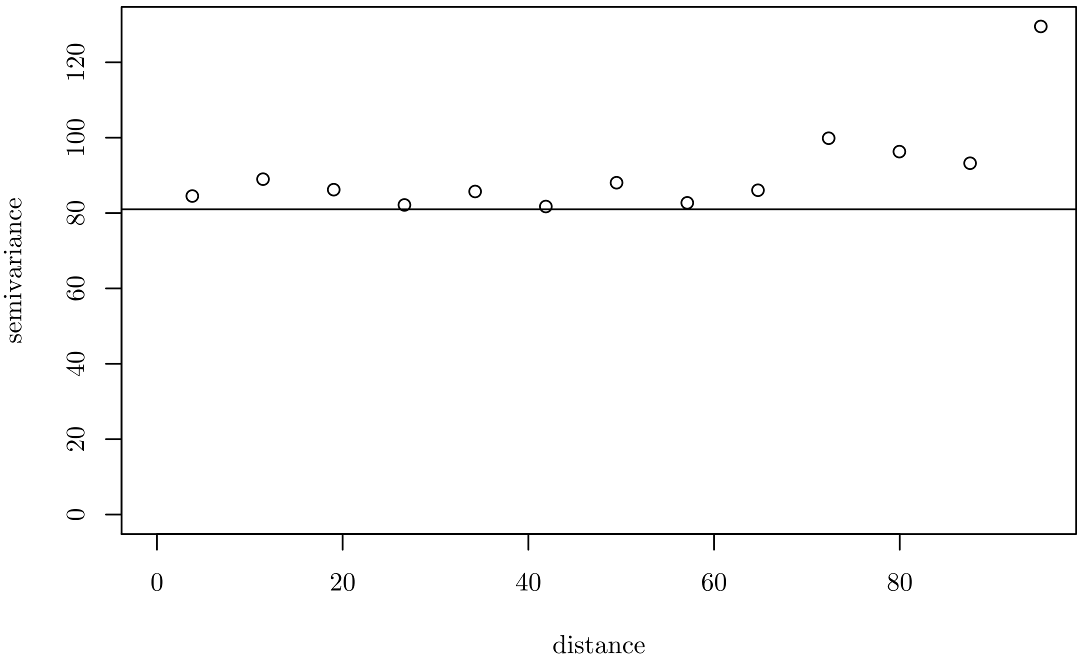
$$\begin{aligned}
E\left[\widehat{\gamma}(\mathbf{h})\right] &= E\left[\frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}\{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\}^2\right] \\
&= E\left[\frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}\{\mu(\mathbf{s}_i) + e(\mathbf{s}_i) - \mu(\mathbf{s}_j) - e(\mathbf{s}_j)\}^2\right] \\
&= E\left[\frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}\left\{[e(\mathbf{s}_i) - e(\mathbf{s}_j)]^2 + [\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)]^2 - [\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j))(e(\mathbf{s}_i) - e(\mathbf{s}_j)]\right\}\right] \\
&= \frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}\left\{E[e(\mathbf{s}_i) - e(\mathbf{s}_j)]^2 + E[\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)]^2 - E[\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j))(e(\mathbf{s}_i) - e(\mathbf{s}_j)]\right\} \\
&= \frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}\left\{E[e(\mathbf{s}_i) - e(\mathbf{s}_j)]^2 + [\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)]^2 - [\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)]E[e(\mathbf{s}_i) - e(\mathbf{s}_j)]\right\} \\
&= \frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}\left\{E[e(\mathbf{s}_i) - e(\mathbf{s}_j)]^2 + [\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)]^2 - [\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)](0)\right\} \\
&= \frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}E[e(\mathbf{s}_i) - e(\mathbf{s}_j)]^2 + \frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}[\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)]^2 \\
&= \gamma_e(\mathbf{h}) + \frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}[\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)]^2
\end{aligned}$$

(c) *Consider the (very) simple model $Z_i = 10 + e_i$ where the e_i are independent normally distributed error terms with variance $\sigma^2 = 81$. We have a pure nugget effect model $\gamma_Z(\mathbf{h}) = \gamma_e(\mathbf{h}) = 81$. Simulate 100 observations of Z_i and calculate the empirical semivariogram assuming the observations are on a one-dimensional transect.*

```
library(geoR)
set.seed(25277)
Zdat <- 10 + rnorm(100, 0, 9)
i <- 1:100
xycoord <- cbind(rep(1, 100), i)
Zvgram <- variog(coords = xycoord, data = Zdat)

variog: computing omnidirectional variogram

par(mar = c(4, 4, 1, 1))
plot(Zvgram)
abline(h = 81)
```



Compare what you see in the plot to the true $\gamma_Z(\mathbf{h})$. Is the result consistent with part (a) above? Why or why not?

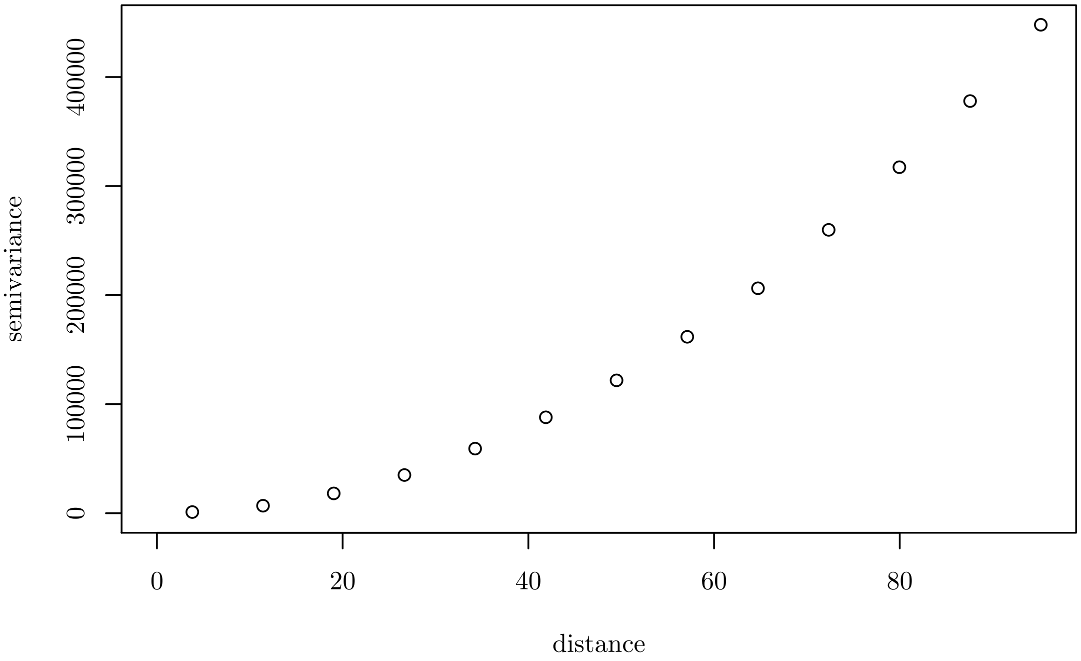
The empirical semivariogram is nearly horizontal and generally stays near the true value of $\gamma(\mathbf{h}) = 81$, which is consistent with the theoretical unbiasedness shown in part (a). The estimates at different lags tend to be higher than the true semivariogram, but that is not surprising because they use the same observed values and are correlated. This is only one observation of a spatial process; Matheron's estimator is unbiased over all realizations of the process.

(d) *Redo the above calculations based on the (still) simple model $Z_i = 10 + 10i + e_i$, i.e. there is now a linear trend and the process is no longer stationary.*

```
Zdat2 <- 10 + 10 * i + rnorm(100, 0, 9)
Zvgram2 <- variog(coords = xycoord, data = Zdat2)

variog: computing omnidirectional variogram

par(mar = c(4, 4, 1, 1))
plot(Zvgram2)
```



Compare the empirical semivariogram `Zvgram2` *to Zvgram. Are the results consistent with part (b) above? Justify your answer.*

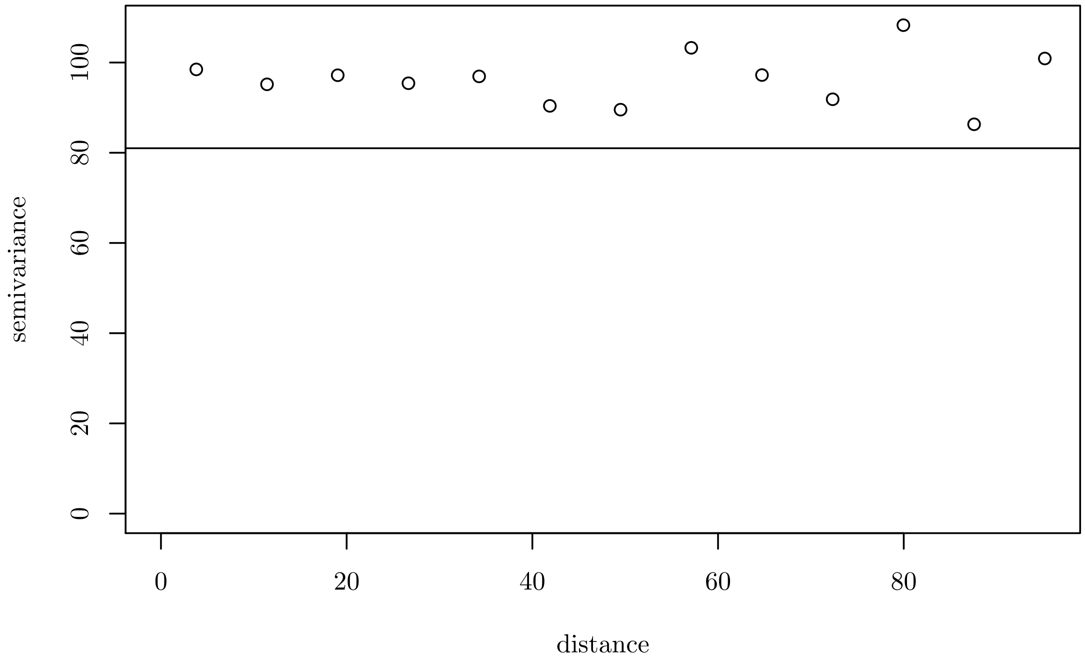
The data in (c) came from a stationary process so that semivariogram is roughly constant. but this empirical semivariogram has a parabolic shape because of son-stationarity. In this situation, the trend in the mean is linear, with $\mu(s_i) - \mu(s_j) = 10(s_i - s_j)$, so the parabolic pattern is consistent with part (b) which shows that the bias in Matheron's estimator is quadratic in the difference in means.

(e) *Fit a linear model to the data in (d), extract the residuals, and compute the empirical semivariogram for the residuals. Note that what you are doing is removing the trend.*

```
e.resid <- residuals(lm(Zdat2 ~ i))
evgram <- variog(coords = xycoord, data = e.resid)
```

```
variog: computing omnidirectional variogram
```

```
par(mar = c(4, 4, 1, 1))
plot(evgram)
abline(h = 81)
```



Compare the 3 empirical semivariograms.

The residuals have a constant mean of zero, so the empirical semivariogram of the residuals does not have a pattern of increasing bias like the one seen in part (d). Like in part (c), the empirical semivariogram is an overestimate but that is probably because of sampling variability (this is another sample of size one, after all).

3. *Attached is a data set containing the carbon nitrogen values used in the carbon/nitrogen data set. The first 2 columns contain the coordinates, total nitrogen is in the third column, total carbon is in the 4th column and the ratio is in the last column. We will work with the total carbon data. Use* `geoR` *for the analysis. It will be easiest if you convert the data into a* `geodata` *object as follows.*

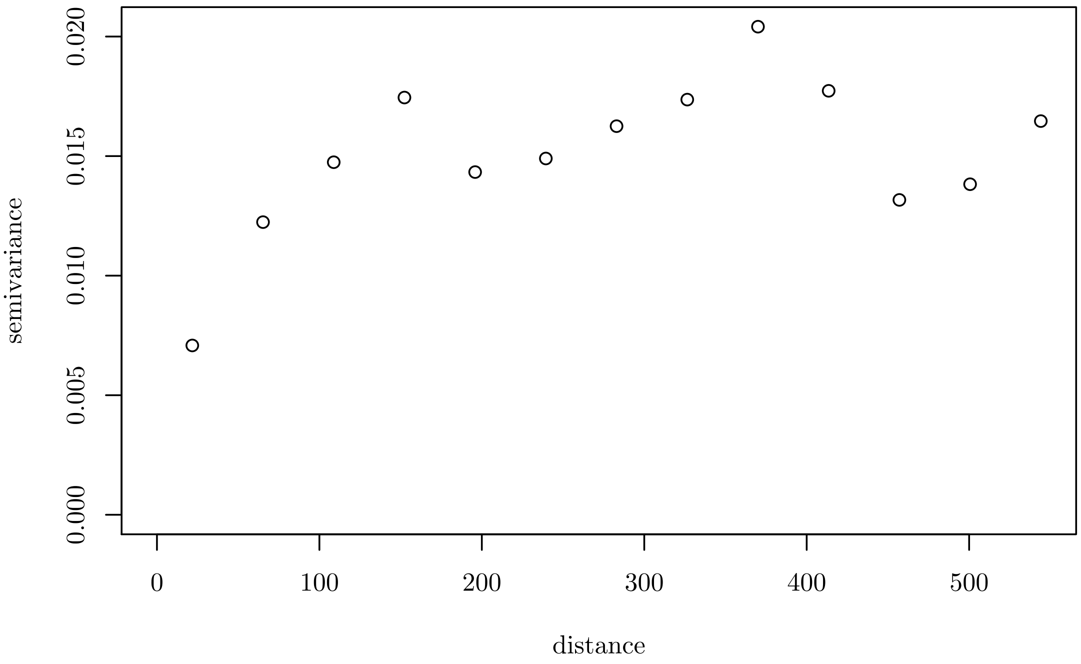
```
CN.dat <- read.table('CN.dat', header = TRUE)
TC.geodata <- as.geodata(CN.dat, coords.col = 1:2, data.col = 4)
```

(a) *Calculate the empirical semivariogram. Give initial eyeball estimates of the nugget effect, sill, and (effective) range.*

```
TC.vgram <- variog(TC.geodata)

variog: computing omnidirectional variogram

par(mar = c(4, 4, 1, 1))
plot(TC.vgram)
```

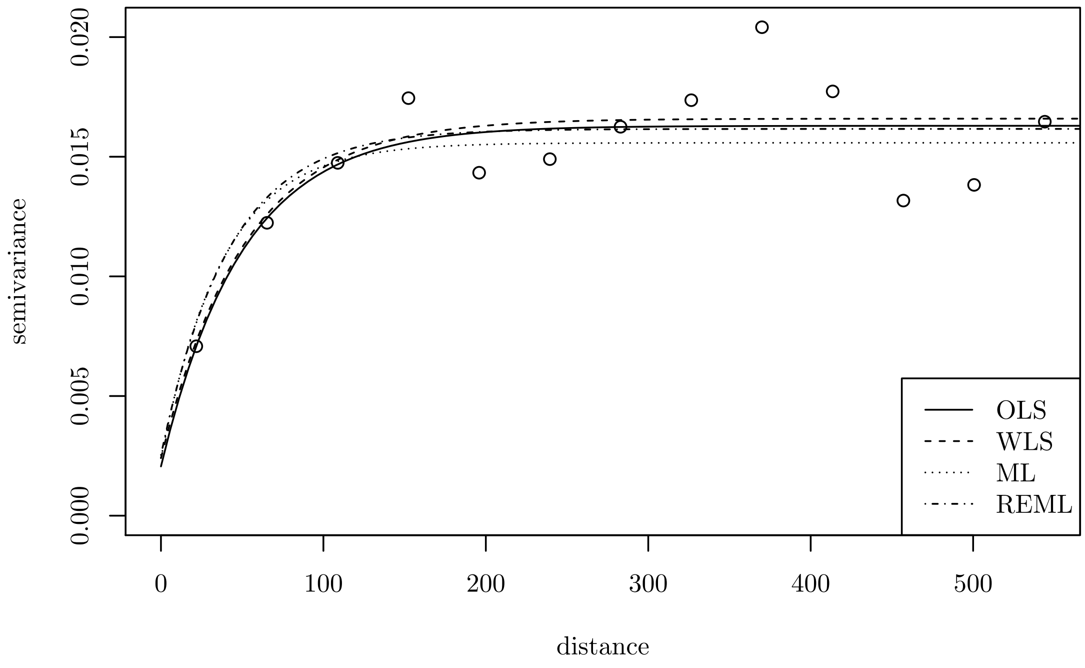


The sill is around 0.015, the effective range looks to be about 150, and the nugget is probably somewhere near 0.005.

(b) *Fit an exponential semivariogram to the carbon data using OLS, WLS, MLE, and REML methods. Specify a nugget effect in each case, i.e. you do not need to consider models without a nugget. Plot the fitted functions and comment on which one you like best.*

```
tc.e.ols <- variofit(TC.vgram, ini.cov.pars = c(0.015, 150/3), nugget = 0.005,
                     fix.nugget = FALSE, cov.model = 'exponential',
                     weights = 'equal')
tc.e.wls <- variofit(TC.vgram, ini.cov.pars = c(0.015, 150/3), nugget = 0.005,
                     fix.nugget = FALSE, cov.model = 'exponential',
                     weights = 'cressie')
tc.e.ml <- likfit(TC.geodata, ini.cov.pars = c(0.015, 150/3), nugget = 0.005,
                  fix.nugget = FALSE, cov.model = 'exponential',
                  lik.method = 'ML')
tc.e.reml <- likfit(TC.geodata, ini.cov.pars = c(0.015, 150/3), nugget = 0.005,
                    fix.nugget = FALSE, cov.model = 'exponential',
                    lik.method = 'RML')

par(mar = c(4, 4, 1, 1))
plot(TC.vgram)
lines(tc.e.ols, lty = 1)
lines(tc.e.wls, lty = 2)
lines(tc.e.ml, lty = 3)
lines(tc.e.reml, lty = 4)
legend('bottomright', lty = 1:4, legend = c('OLS', 'WLS', 'ML', 'REML'))
```

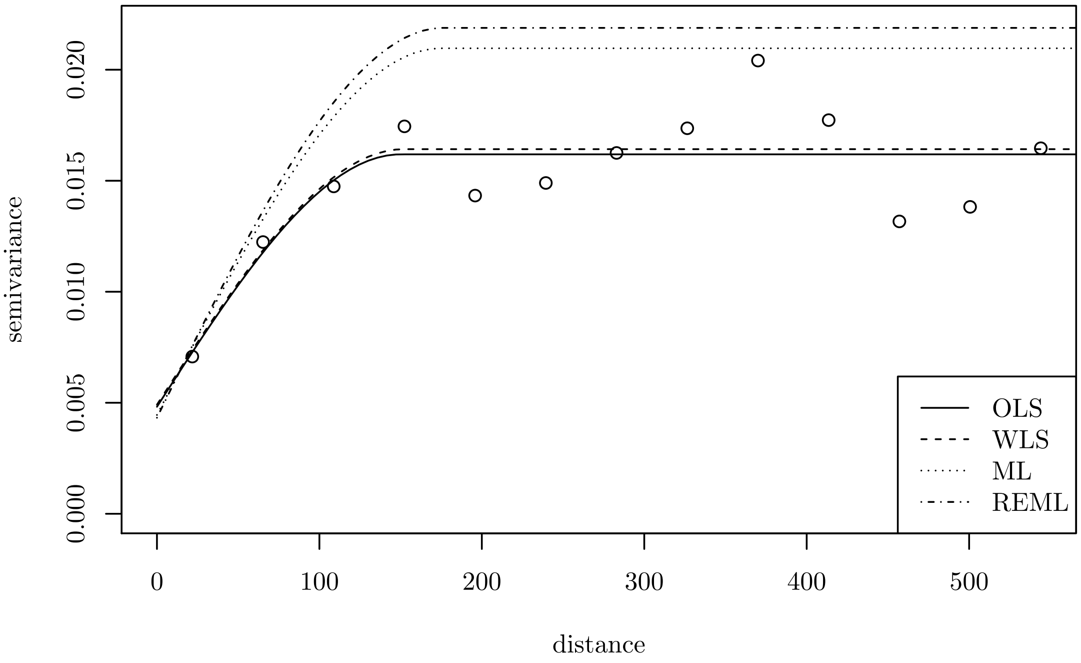


All of the estimates are similar, but the OLS and WLS estimates look best because they come closest to $\widehat{\gamma}$ at the smallest lags. I would use WLS just because I like the theoretical idea of downweighting the large lags even though the result is barely distinguishable from the OLS result here.

(c) *Redo part (b) by fitting a spherical model.*

```
tc.s.ols <- variofit(TC.vgram, ini.cov.pars = c(0.015, 150), nugget = 0.005,
                     fix.nugget = FALSE, cov.model = 'spherical',
                     weights = 'equal')
tc.s.wls <- variofit(TC.vgram, ini.cov.pars = c(0.015, 150), nugget = 0.005,
                     fix.nugget = FALSE, cov.model = 'spherical',
                     weights = 'cressie')
tc.s.ml <- likfit(TC.geodata, ini.cov.pars = c(0.015, 150), nugget = 0.005,
                  fix.nugget = FALSE, cov.model = 'spherical',
                  lik.method = 'ML')
tc.s.reml <- likfit(TC.geodata, ini.cov.pars = c(0.015, 150), nugget = 0.005,
                    fix.nugget = FALSE, cov.model = 'spherical',
                    lik.method = 'RML')

par(mar = c(4, 4, 1, 1))
plot(TC.vgram, ylim = c(0, 0.022))
lines(tc.s.ols, lty = 1)
lines(tc.s.wls, lty = 2)
lines(tc.s.ml, lty = 3)
lines(tc.s.reml, lty = 4)
legend('bottomright', lty = 1:4, legend = c('OLS', 'WLS', 'ML', 'REML'))
```



Again, I would choose WLS. This time, ML and REML both overestimate the sill, but WLS and OLS match $\widehat{\gamma}$ pretty well.

(d) *Summarize your results in table format. Compare the results and discuss.*

```
compare <- do.call(rbind,
                   lapply(list(tc.e.ols, tc.e.wls, tc.e.ml, tc.e.reml,
                               tc.s.ols, tc.s.wls, tc.s.ml, tc.s.reml),
                          function(x){
                            return(data.frame(
                                     Model = x$cov.model,
                                     Method = x$method,
                                     Nugget = x$nugget,
                                     Sill = x$cov.pars[1],
                                     Range = x$cov.pars[2] *
                                       ifelse(x$cov.model == 'exponential', 3, 1)))
                          }))
xtable(compare, digits = 5)
```

Model	Method	Nugget	Sill	Range
exponential	OLS	0.00206	0.01423	150.00000
exponential	WLS	0.00241	0.01418	154.62223
exponential	ML	0.00241	0.01317	115.01824
exponential	RML	0.00254	0.01362	125.47982
spherical	OLS	0.00482	0.01137	150.00000
spherical	WLS	0.00491	0.01151	152.03148
spherical	ML	0.00445	0.01651	174.82292
spherical	RML	0.00433	0.01756	176.06548

I multiplied the exponential range parameter by three so that the table actually shows the effective range. I tried a few different initial range values, and as long as the initial range was somewhere in the flat section of the semivariogram, the OLS and WLS range estimate was close to the initial value (for both models). For a given model, all four estimation methods gave about the same nugget estimate, though the nugget estimates for the spherical model were almost 2.5 times as big as the nugget estimates for the exponential model. OLS and WLS gave similar sill values, and ML and REML also found similar sill estimates; for the exponential model all four methods found pretty similar sills, but for the spherical model OLS/WLS and ML/REML yielded noticeably different sill values. The conclusion I can make from all this is that fitting a semivariogram model involves many choices that affect the end result, so you need to check the fitted model against the empirical semivariogram to see if it looks reasonable. For this dataset, I would use the exponential model fit by WLS.