STA 5377, Homework 2

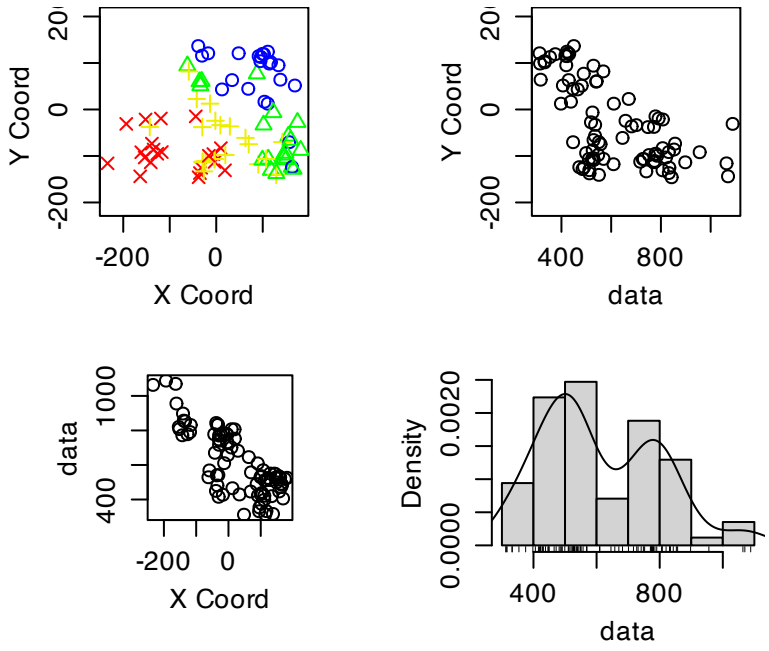
Carson Slater *Baylor University*

1

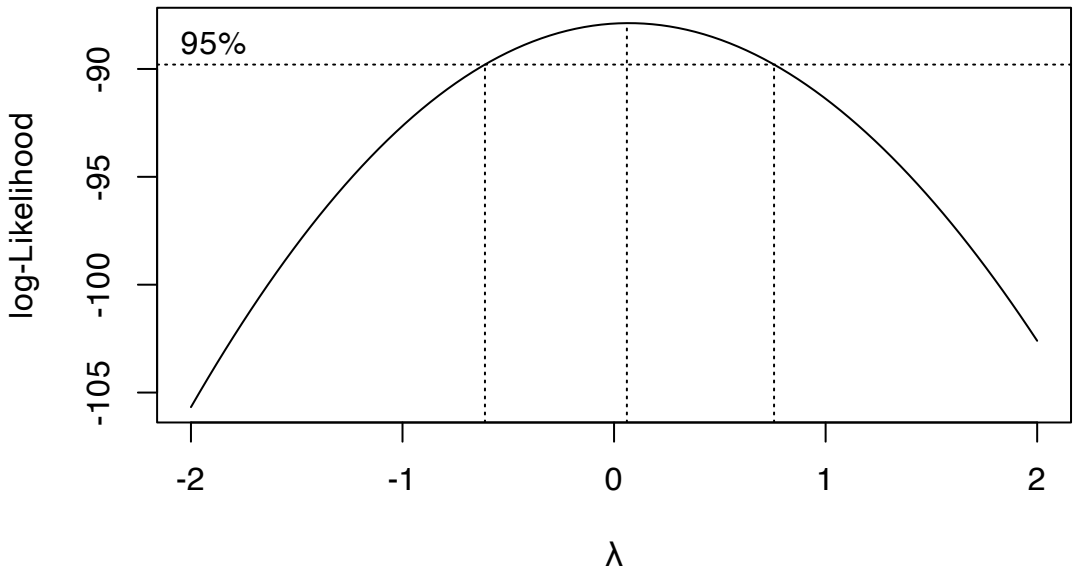
The `wolfcamp` data set is included in the package `geoR`. The data consist of the locations of 85 wells in the Wolfcamp aquifer in Texas. For each well, the piezometric head was measured; essentially, this is the level of the water table above sea level. `wolfcamp$coord` contains the locations of the observations; `wolfcamp$data` contains the Z values. The structure of this problem is a little different from what we have seen before. Provided below is some R code that will allow you to conduct a few analyses of these data. You are to execute this code and for each line explain what it is doing. You will probably want to make reference to the manual for `geoR` as you do this. Finally, based on the output from the code, you are to summarize your findings for these data.

(a)

Perform exploratory spatial data analysis.



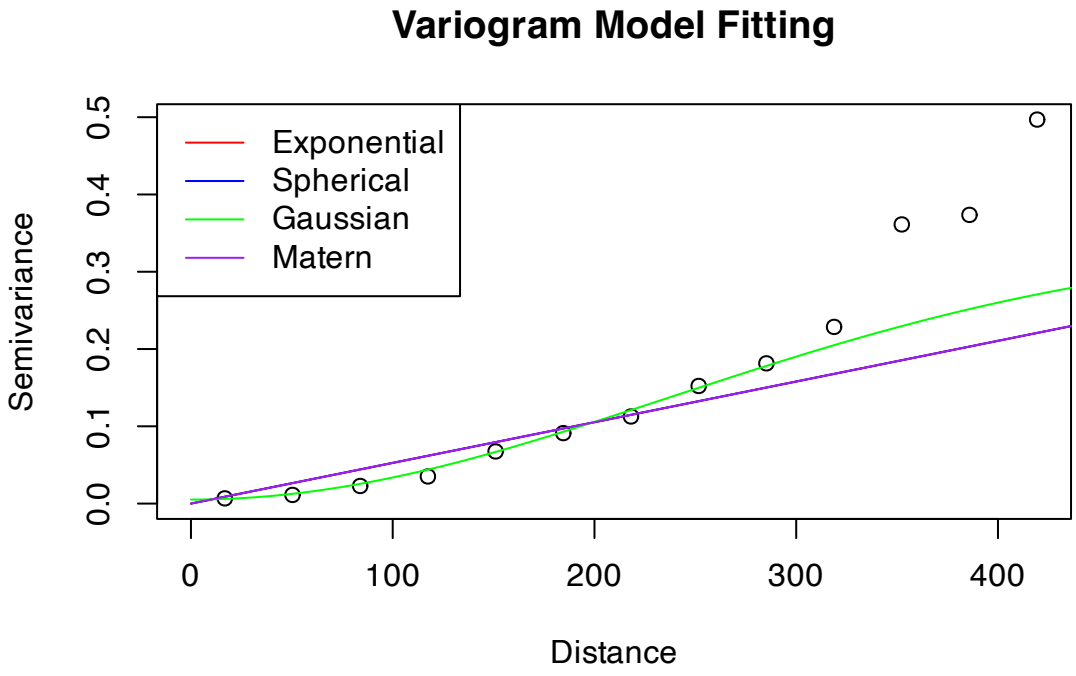
For the spatial plot (top right), there appears to be clustering for binned measurments of z values. Both the x and y coordinates seem to be negatively correlated with the z values. Additionally, the z values appear to have a bimodal distribution.



Per this Box-Cox graphic, it appears that these data would exhibit normality after a log-transformation. We elect to transform the data, as our work in part (d) works only under the normality assumption.

(b)

Compute and plot the empirical semivariograms.



(c)

Consider four theoretical variogram models, exponential, spherical, Gaussian, and matern for variogram fitting using WLS. Find the best fit.

	Exponential	Spherical	Gaussian	Matern
WSS	395.8723	395.8729	63.31429	395.8723

Here it appears the Gaussian covariance structure yielded the lowest weighted sum of squares, giving evidence that it is the most viable model.

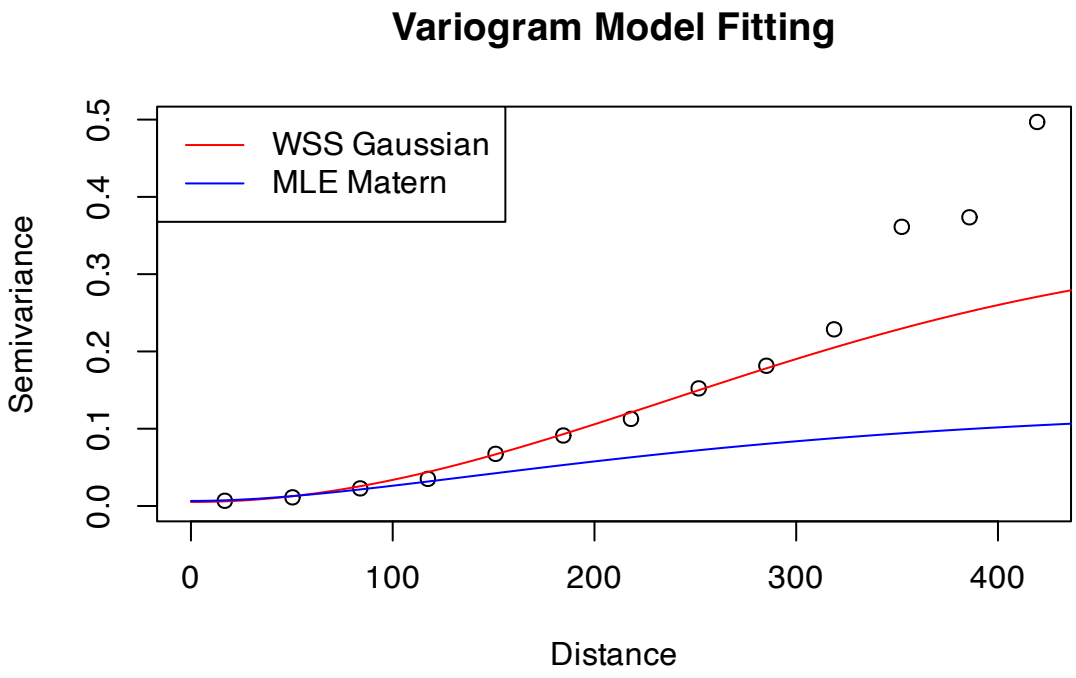
(d)

Table 2: Performance of each covariance structure using the MLE for estimation.

Model	LogLikelihood	AIC
Exponential	68.45128	-128.9026
Spherical	69.30363	-130.6073
Gaussian	68.23107	-128.4621
Matern	69.36614	-130.7323

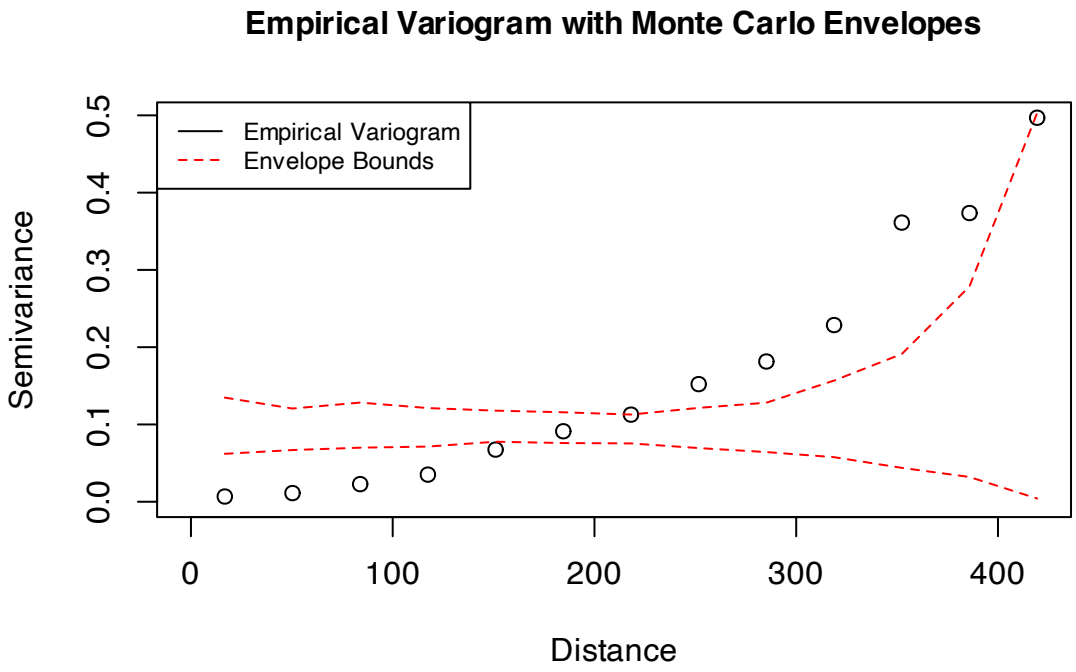
From the table, the best model was the Matern model. The Gaussian model was unable to be fit due to a singular matrix issue.

(e)



It appears the empirical semivariogram with the gaussian covariance model is the best by far.

(f)



These envelopes for the semivariogram estimates seem to indicate that there is likely substantial spatial

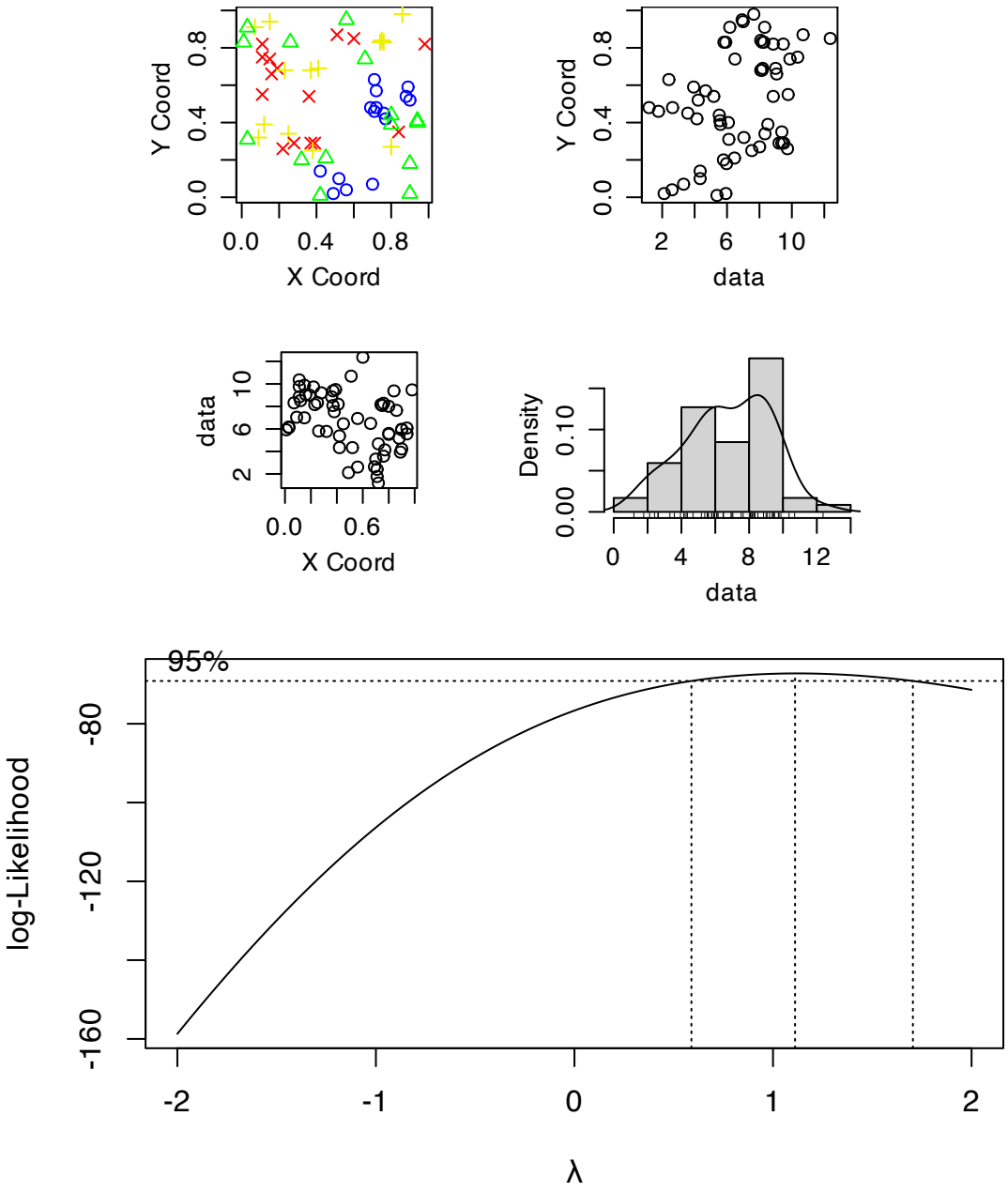
correlation at up to $s = 150$ units).

2

Consider the `wells` data (`wells.txt`). `x` and `y` are spatial coordinates, and `z` is the variable of interest.

(a)

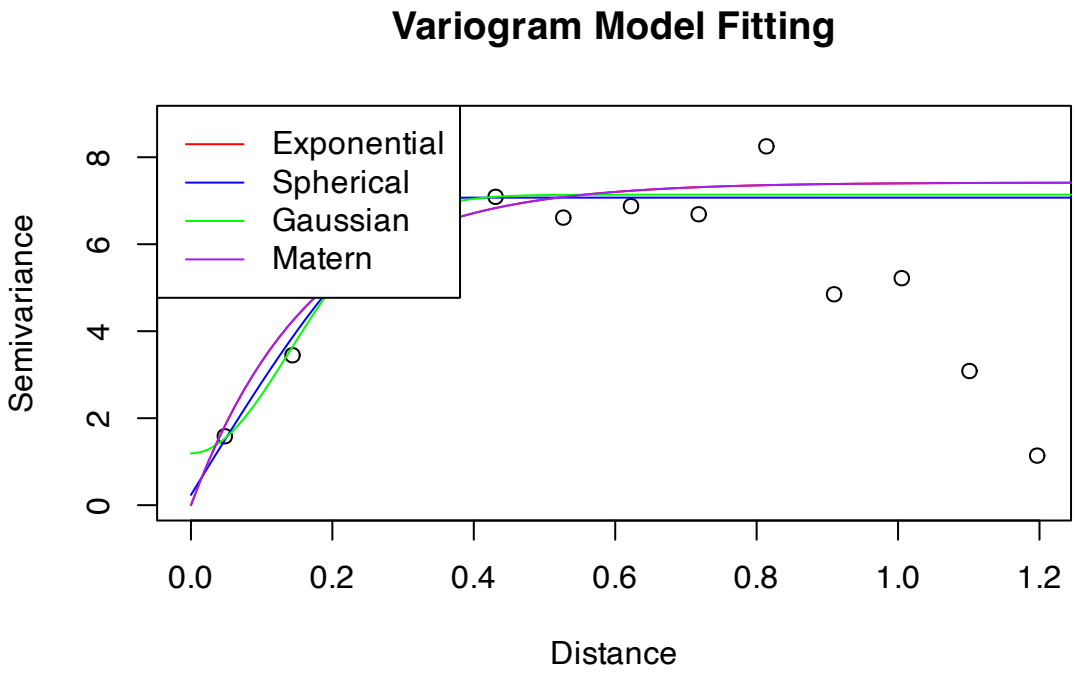
Perform exploratory spatial data analysis.



This data seems to have clusters of the ranges in the top left and bottom right of the data plot (see red "x" points and blue "o" points). There does not seem to be any substantial correlation between the x coordinates and the data, and the y coordinates and the data. These data do not seem to need a transformation.

(b)

Compute and plot the empirical semivariograms.



(c)

Consider four theoretical varigoram models, exponential, spherical, Gaussian, and matern for variogram fitting using WLS. Find the best fit.

	Exponential	Spherical	Gaussian	Matern
WSS	56.68289	40.01399	40.87849	56.68289

Here it appears the Spherical covariance structure yielded the lowest weighted sum of squares, giving evidence that it is the most viable model.

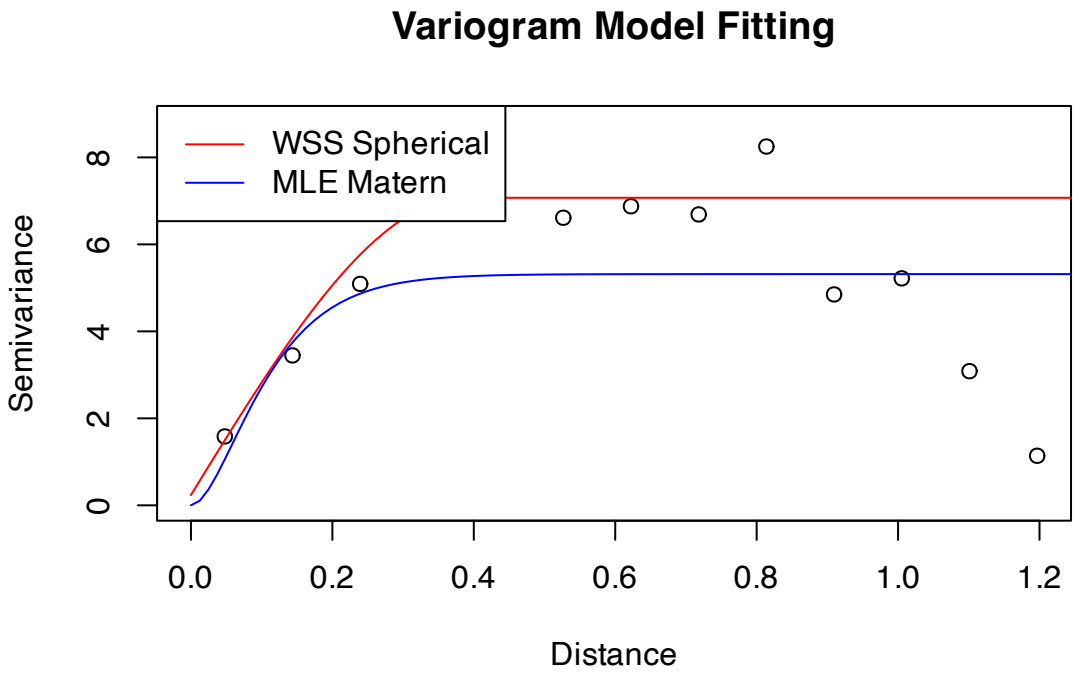
(d)

Table 4: Performance of each covariance structure using the MLE for estimation.

Model	LogLikelihood	AIC
Exponential	-111.2289	230.4578
Spherical	-110.2738	228.5476
Gaussian	-110.7989	229.5978
Matern	-108.5703	225.1406

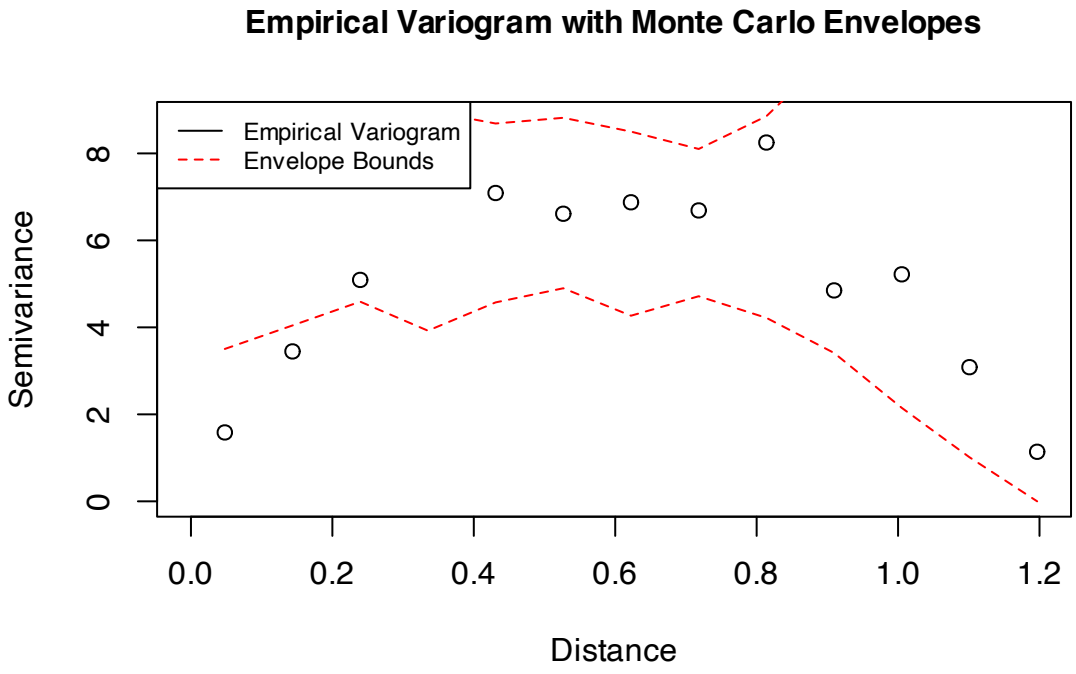
From the table, the best model was the Matern model, with the highest log-likelihood and the lowest AIC.

(e)



It appears the MLE semivariogram was actually better than the WLS Spherical semivariogram. This is debatable, as the MLE could be underestimating the partial sill, but the empirical semivariogram also tapers down substantially.

(f)



These envelopes for the semivariogram estimates seem to indicate that there may be spatial correlation at very small distances ($s \leq 0.2$ units).

3

The variance of the empirical variogram $\hat{\gamma}(h)$ is given by:

$$\text{Var}(2\hat{\gamma}(h)) \approx \frac{2(2\gamma(h))^2}{N(h)}.$$

Since we usually do not know the true $\gamma(h)$, we can plug in an estimate:

$$\text{Var}(2\hat{\gamma}(h)) \approx \frac{2(2\hat{\gamma}(h))^2}{N(h)}.$$

Rewriting in terms of the semivariogram:

$$\text{Var}(\hat{\gamma}(h)) \approx \frac{(2\hat{\gamma}(h))^2}{2N(h)}.$$

Thus, the standard deviation of $\hat{\gamma}(h)$ is approximately:

$$\frac{\sqrt{2\hat{\gamma}(h)}}{\sqrt{N(h)}}.$$

An approximate 95% confidence interval for $\gamma(h)$ is:

$$\gamma(h) \in \hat{\gamma}(h) \pm 1.96 \frac{\sqrt{2\hat{\gamma}(h)}}{\sqrt{N(h)}}.$$

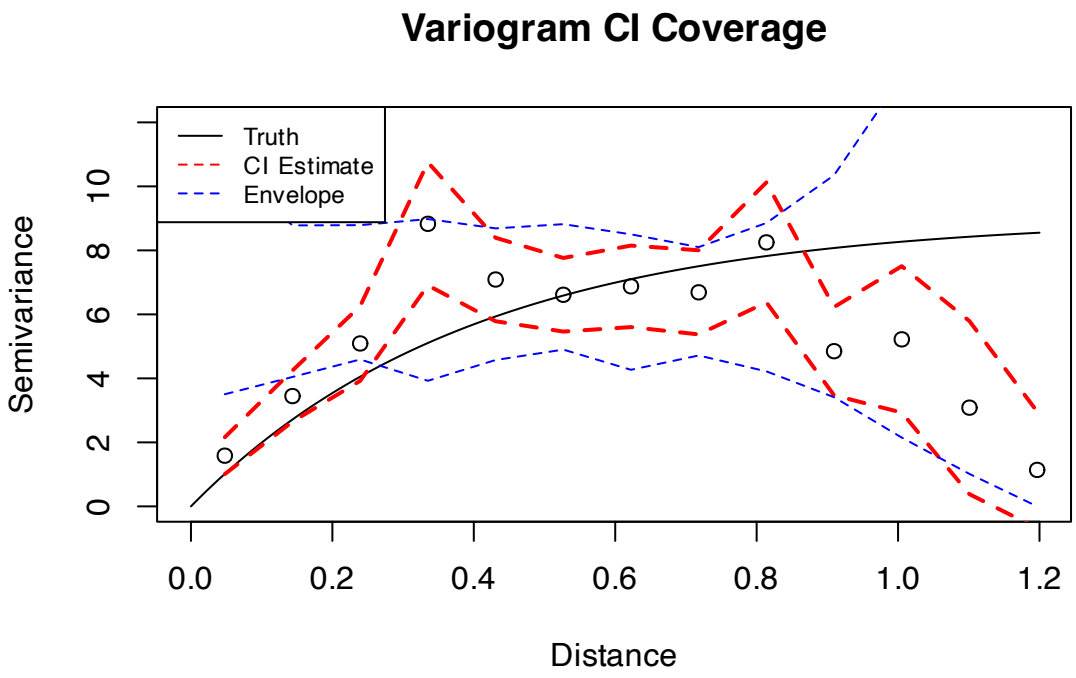
Assume that the true variogram is exponential with $c_0 = 0$, $c_1 = 9$, and $\phi = 0.4$. How well do the confidence intervals cover the true semivariogram? Briefly summarize your findings and compare with the result of `vario.mc.env`.

The exponential semivariogram is written as

$$\gamma(h) = c_0 + c_1 \left(1 - e^{-h/\phi}\right) = 9 \left(1 - e^{-h/0.4}\right).$$

Therefore if we take the true variogram and find where the points are, say we have $N(h) = 13$ bins of distances, and we estimate $\hat{\gamma}(h)$ using the `variog()` function. The bin limits for the empirical variogram are shown below:

0.0000000	0.1914751	0.3829502	0.5744254	0.7659005	0.9573756	1.148851
0.0957376	0.2872127	0.4786878	0.6701629	0.8616380	1.0531132	1.244588



Confidence Interval	Envelope
0.6153846	0.7692308

It seems for the confidence intervals, the coverage probability is around 61% and for the envelope, it is around 77%.

Appendix for Problem 3

For question 3, I misunderstood the question initially and wrote a small simulation to determine roughly what this coverage probability is. Here is the simulation with the results:

```
# True variogram parameters
c0 <- 0
c1 <- 9
phi <- 0.4
n_sim <- 1000  # Number of simulations
n_points <- 100  # Number of spatial locations

set.seed(613)

# Function to generate data and compute empirical variogram
simulate_variogram <- function() {
  # Generate random spatial coordinates
  coords <- cbind(runif(n_points, 0, 10), runif(n_points, 0, 10))

  # Simulate spatial data from an exponential variogram
  geodata <- grf(n = n_points, grid = coords, cov.pars = c(c1, phi), nugget =

  # Compute empirical variogram
  emp_vario <- variog(geodata, messages = FALSE)

  # Compute approximate confidence intervals
  var_est <- (2 * emp_vario$v)2 / (2 * emp_vario$n)
  ci_lower <- emp_vario$v - 1.96 * sqrt(var_est)
  ci_upper <- emp_vario$v + 1.96 * sqrt(var_est)

  # Compute true variogram values
  true_gamma <- c1 * (1 - exp(-emp_vario$u / phi))

  # Check if true variogram is within confidence intervals
  ci_coverage <- mean(true_gamma >= ci_lower & true_gamma <= ci_upper)

  # Compute permutation-based confidence envelopes using vario.mc.env
  mc_env <- geoR::variog.mc.env(geodata, obj.variog = emp_vario, nsim = 100,
```

```
  mc_env_coverage <- mean(mc_env$v.lower <= true_gamma & true_gamma <= mc_env

  c(ci_coverage, mc_env_coverage)
}

# Run simulations in parallel
coverage_results <- future_map(1:n_sim, ~simulate_variogram(), .progress = TR

# Convert results to dataframe
coverage_df <- bind_rows(lapply(coverage_results, \(x) tibble(ci_coverage = x
```

Running the simulation above, we have the following results.

Mean CI Coverage	Mean Envelope Coverage
0.5820321	0.7897949

These simulation numbers are close to the numbers we found in **3**.