HW5- Spatial Models

Sarah Wiegreffe

February 17, 2016

```r
library(vegan)
library(nlme)
data(BCI)
## UTM Coordinates (in metres)
BCI_xy = data.frame(x = rep(seq(625754, 626654, by=100), each=5),
                    y = rep(seq(1011569,  1011969, by=100), len=50))
```

1) Examine if there is evidence of spatial dependence in a rare and a common species in the BCI tree dataset

```r
#subset dataset on rare and common species
sub_common = BCI$Faramea.occidentalis
sub_rare= BCI$Heisteria.concinna

#check that models have few datapoints with 0 abundance and change species as necessary
#hist(sub_common)
#hist(sub_rare)
```

```r
col_brks = hist(sub_common, plot=F)$breaks
col_indices = as.numeric(cut(sub_common, col_brks))
cols = rev(terrain.colors(length(col_brks)))
plot(BCI_xy, cex=2, pch=19, col=cols[col_indices])
```
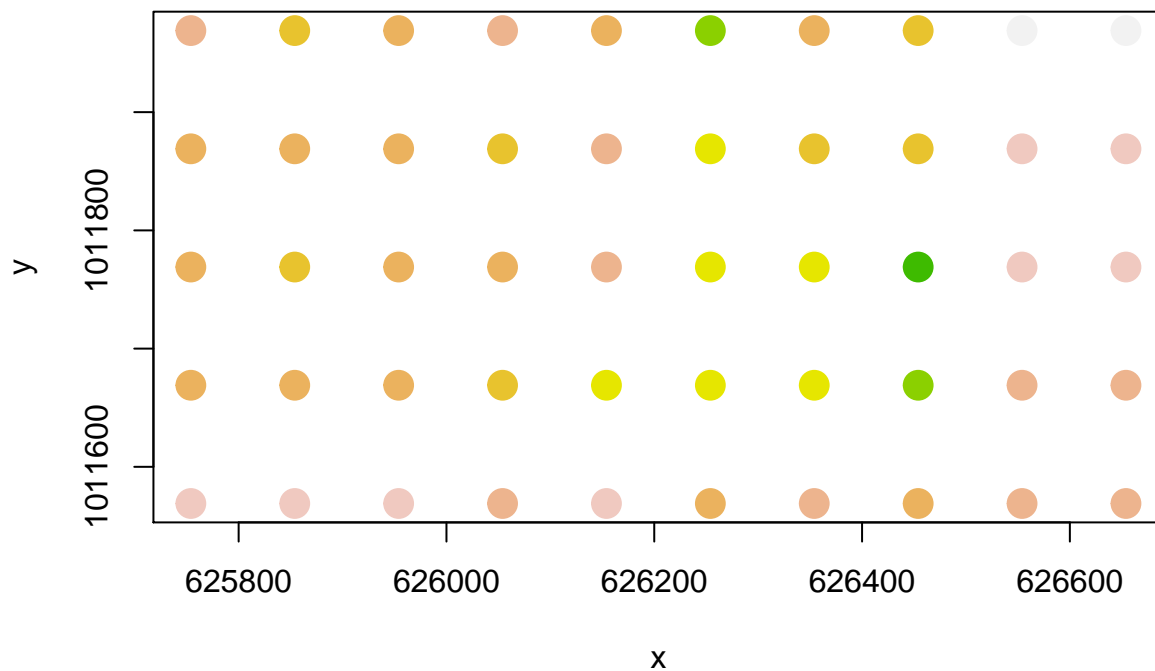
```
col_brks = hist(sub_rare, plot=F)$breaks
col_indices = as.numeric(cut(sub_rare, col_brks))
cols = rev(terrain.colors(length(col_brks)))
plot(BCI_xy, cex=2, pch=19, col=cols[col_indices])
```
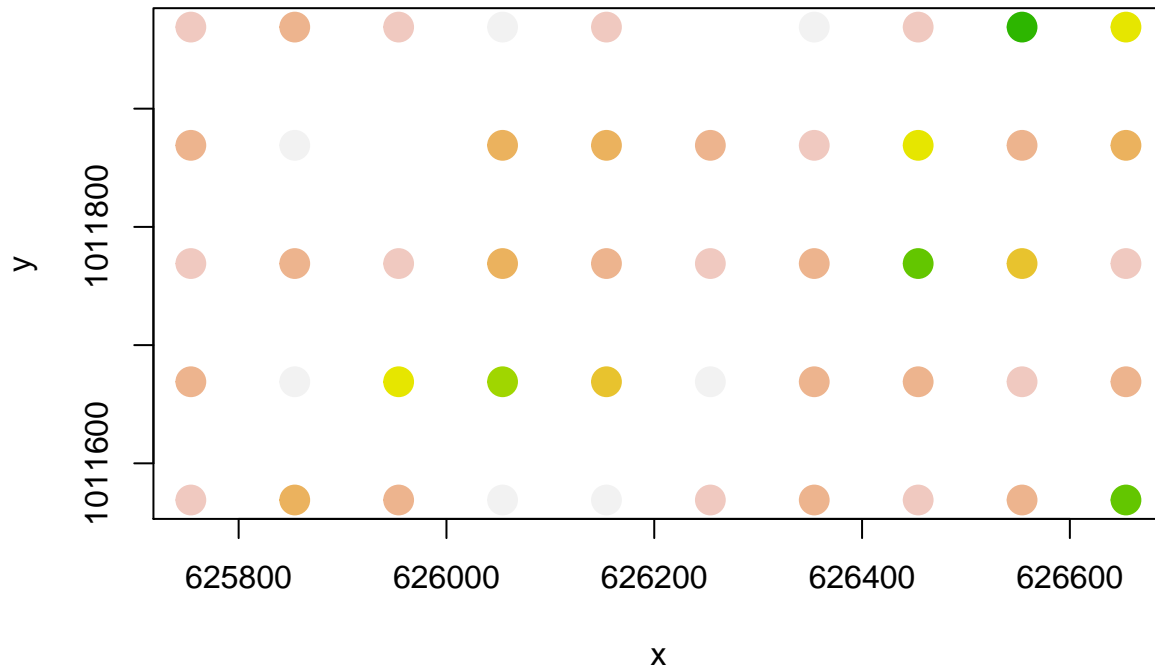


```
# calculate Euclidean distance between abundance and spatial coordinates
abundance_dist_common = dist(sub_common)
abundance_dist_rare = dist(sub_rare)
xy_dist = dist(BCI_xy)

max_dist = max(xy_dist) / 2

# plot results
plot(xy_dist, abundance_dist_common)
abline(lm(abundance_dist_common ~ xy_dist), lwd=3, col='red')
lines(lowess(xy_dist, abundance_dist_common), lwd=3, col='pink')
abline(v = max_dist, col='red', lwd=3, lty=2)
```
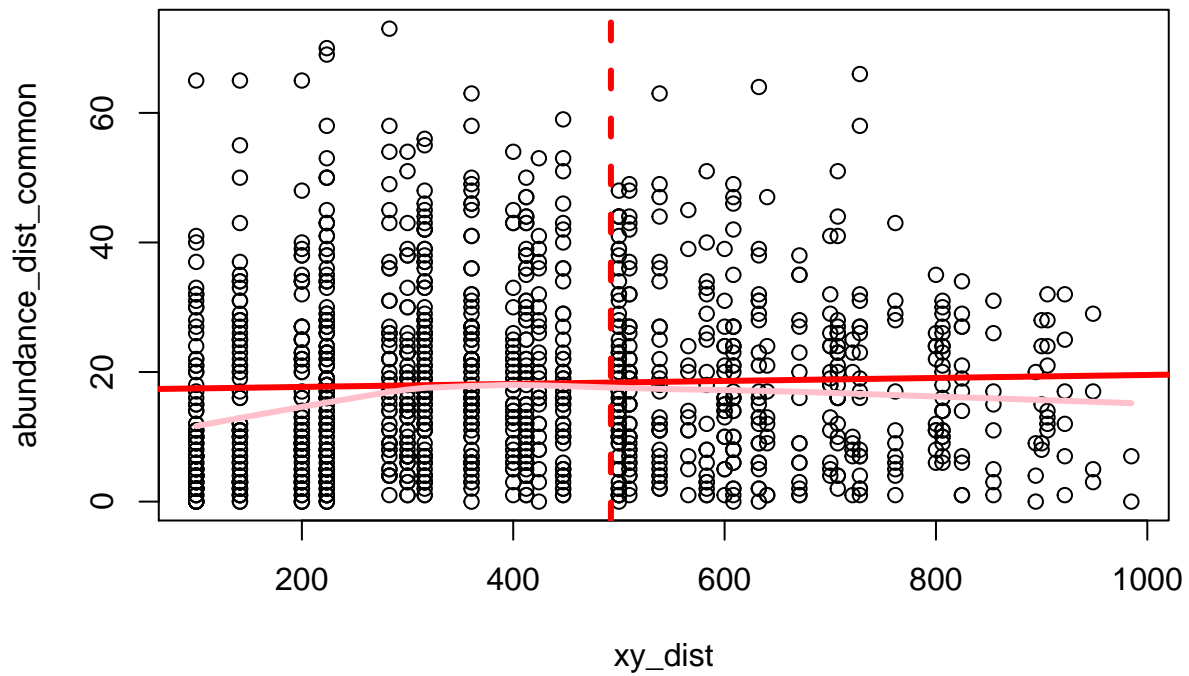
```r
plot(xy_dist, abundance_dist_rare)
abline(lm(abundance_dist_rare ~ xy_dist), lwd=3, col='red')
lines(lowess(xy_dist, abundance_dist_rare), lwd=3, col='pink')
abline(v = max_dist, col='red', lwd=3, lty=2)
```
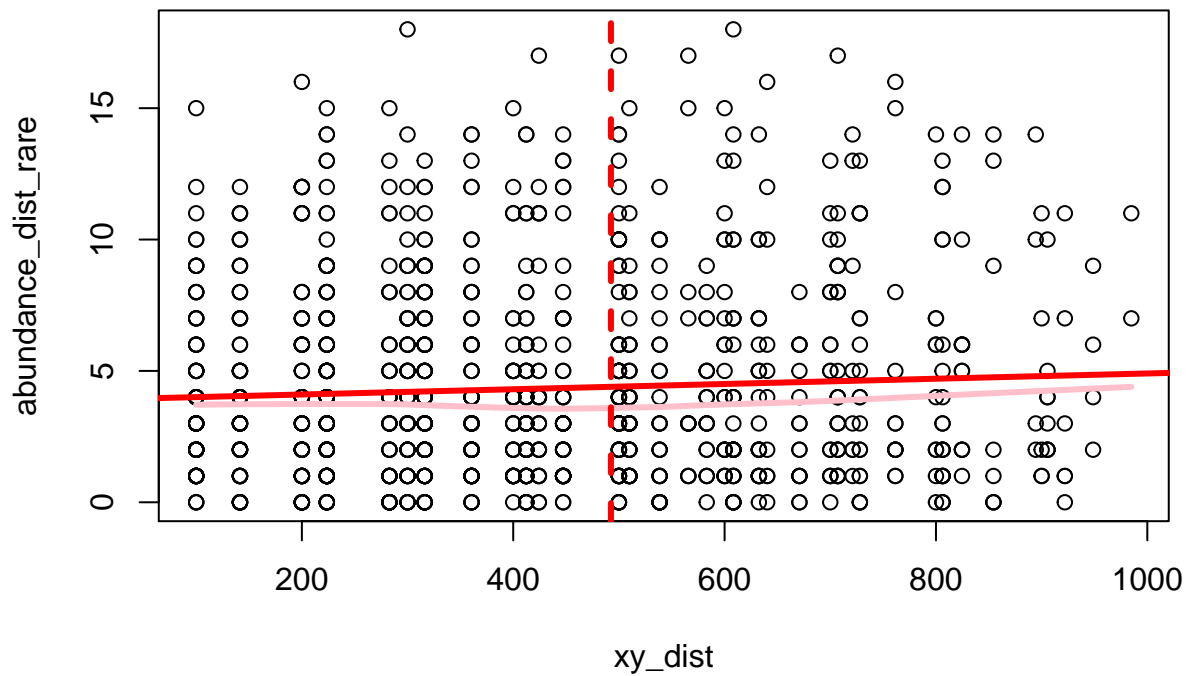


```r
# compute correlation
obs_cor_rare = cor(xy_dist, abundance_dist_rare)
obs_cor_common = cor(xy_dist, abundance_dist_common)
obs_cor_rare
```

```
## [1] 0.05686859
```

```
obs_cor_common
```

```
## [1] 0.0355889
```

```
# carry out a permutation test for significance:
nperm = 1000
null_cor = obs_cor_rare
for (i in 2:nperm) {
    # shuffle the rows of the spatial coordinates
    tmp_xy = BCI_xy[sample(nrow(BCI_xy)), ]
    # correlation between the shuffled spatial coordinates and sr_dist
    null_cor[i] = cor(dist(tmp_xy), abundance_dist_rare)
}
# compute the p-value
sum(null_cor >= obs_cor_rare) / nperm
```

```
## [1] 0.165
```

```
nperm = 1000
null_cor = obs_cor_common
for (i in 2:nperm) {
    # shuffle the rows of the spatial coordinates
    tmp_xy = BCI_xy[sample(nrow(BCI_xy)), ]
    # correlation between the shuffled spatial coordinates and sr_dist
    null_cor[i] = cor(dist(tmp_xy), abundance_dist_common)
}
# compute the p-value
sum(null_cor >= obs_cor_common) / nperm
```

```
## [1] 0.255
```

I chose to look at Heisteria.concinna as my rare species and the Faramea.occidentalis species for my common ones. From the visualization of locations of samples colored by their abundance for both species, there does not seem to be evidence of spatial dependence judging by the distribution of colors (abundances) over the plot. In the second plots, both hae a line slope of zero which implies there is no correlation between xy-position and abundance. Correlation values are very close to zero which supports this. Also, neither species had a p-value that showed significance of spatial dependence different from that of random chance from a permutation test.

2) Build two generalized linear models to predict the abundance of the species *Drypetes standleyi* using the abundance of other tree species in the study site. Specifically examine the following species as predictor variables.

- Model 1: only include a single species as a predictor variable

- Model 2: include all of the species as predictor variables

With both models examine the spatial dependence of the residuals using the function `Variogram`. Model the spatial dependence in the residuals using one of the error structures available.

- Did including the spatial error term have a large impact on the coefficients of the model?

- Did including the spatial error terms significantly improve model fit (use function `anova` to carry out model comparison)?

- Explain why modeling the spatial errors has a large influence in the first model but not in the second model.

```r
sp_ids = c("Cordia.lasiocalyx", "Hirtella.triandra",
           "Picramnia.latifolia", "Quassia.amara",
           "Tabernaemontana.arborea", "Trattinnickia.aspera",
           "Xylopia.macrantha")

abundance_response = BCI$Drypetes.standleyi

abundance_df = data.frame(BCI$Drypetes.standleyi,BCI[,which(colnames(BCI) %in% sp_ids)],BCI_xy)

#model with single predictor variable- Model 1
abundance_lm_single = gls(abundance_response ~ Cordia.lasiocalyx, data=abundance_df,
        corr=corExp(form=~ x + y), control = glsControl(opt='optim',  msVerbose=T))

par(mfrow=c(1,1))
plot(Variogram(abundance_lm_single, form= ~ x + y))
```
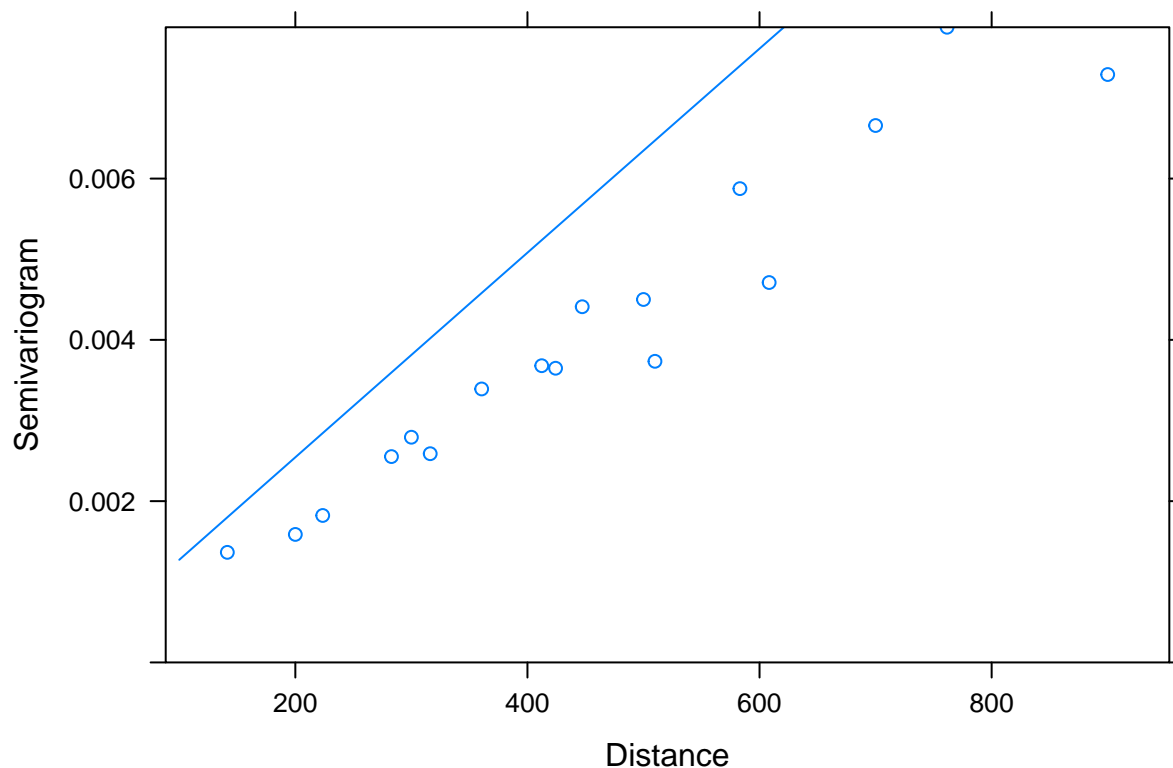


```r
abundance_exp_single = update(abundance_lm_single, corr=corExp(form=~x + y))
# examine fit of error model to the raw model residuals
#plot(Variogram(abundance_exp_single, maxDist = max_dist))
# examine the normalized residuals
#plot(Variogram(abundance_exp_single, resType='normalized', maxDist = max_dist))
```

5

```
# let's look at the same model but with a nugget
abundance_exp_nug_single = update(abundance_exp_single, corr=corExp(form=~x + y, nugget=T))
# Look to provide a better fit:
#plot(Variogram(abundance_exp_nug_single, maxDist = max_dist))
#plot(Variogram(abundance_exp_nug_single, resType = 'n', maxDist = max_dist))

# let's examine the rational quadratic error model
abundance_rat_nug_single = update(abundance_lm_single, corr=corRatio(form=~x + y, nugget=T))

# examine fit of error model to model residuals
plot(Variogram(abundance_rat_nug_single, maxDist = max_dist))
```
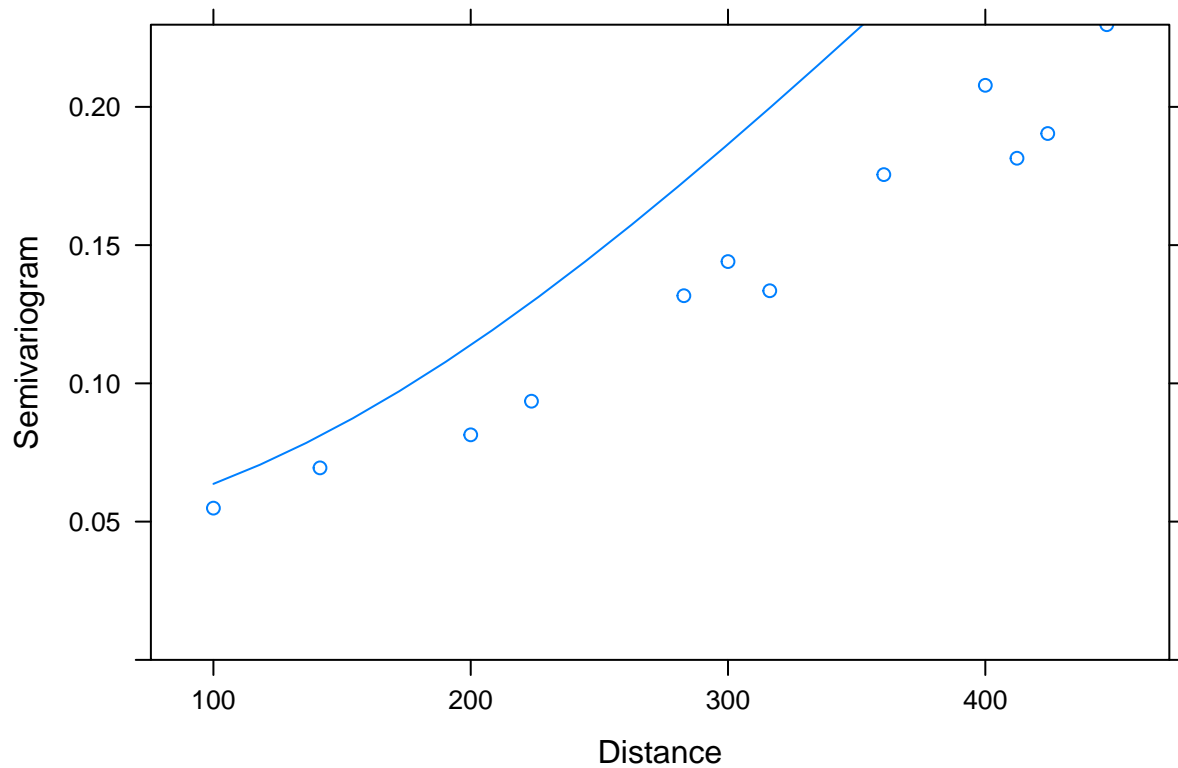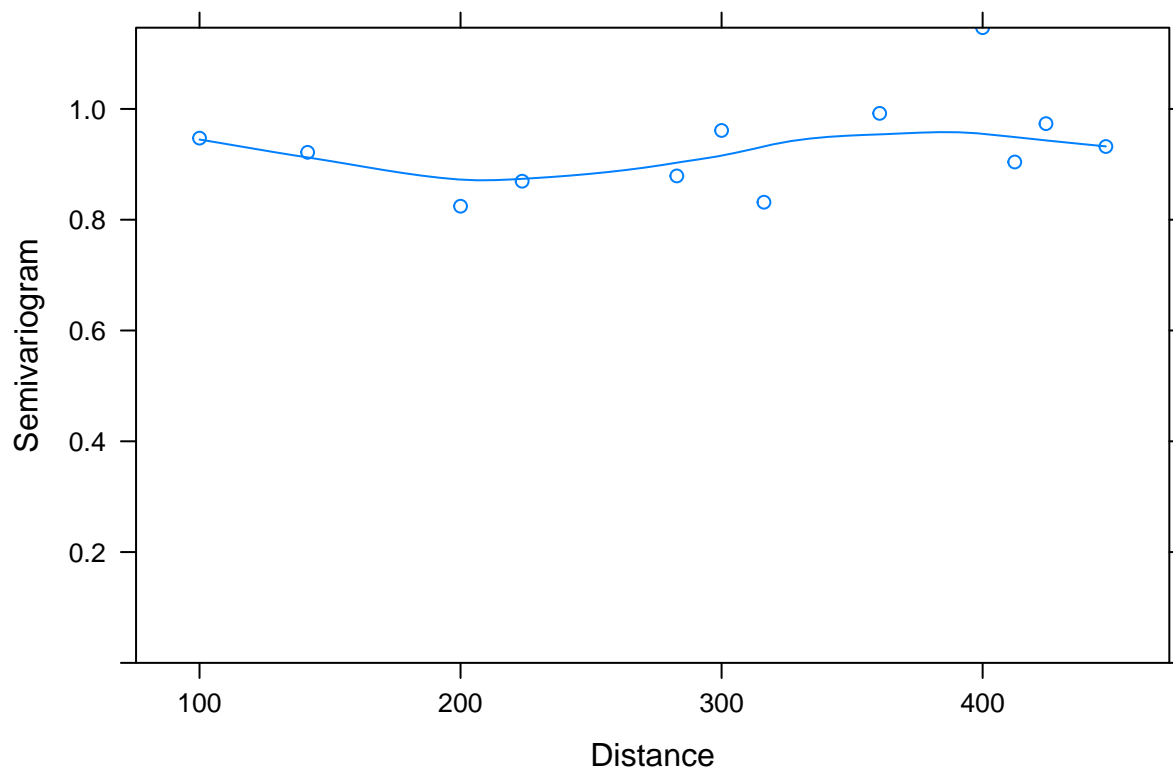


```
# Good fit:
plot(Variogram(abundance_rat_nug_single, resType='n', maxDist = max_dist))
```

```
# let's compare the models
anova(abundance_lm_single, abundance_exp_single, abundance_exp_nug_single,
      abundance_rat_nug_single, test=F)
```

```
##                          Model df      AIC      BIC     logLik
## abundance_lm_single          1  4 310.6456 318.1304 -151.3228
## abundance_exp_single         2  4 310.6456 318.1304 -151.3228
## abundance_exp_nug_single     3  5 311.0600 320.4160 -150.5300
## abundance_rat_nug_single     4  5 308.6589 318.0149 -149.3295
```

```
summary(abundance_rat_nug_single)
```

```
## Generalized least squares fit by REML
##   Model: abundance_response ~ Cordia.lasiocalyx
##   Data: abundance_df
##        AIC      BIC    logLik
##   308.6589 318.0149 -149.3295
##
## Correlation Structure: Rational quadratic spatial correlation
##  Formula: ~x + y
##  Parameter estimate(s):
##        range        nugget
## 720.94330740    0.04563307
##
## Coefficients:
##                      Value Std.Error   t-value p-value
## (Intercept)      12.861472 13.982372 0.9198348  0.3623
## Cordia.lasiocalyx 0.104806  0.211538 0.4954484  0.6225
```

7

```
##
##  Correlation:
##                 (Intr)
## Cordia.lasiocalyx -0.157
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -0.7674686 -0.7172480 -0.6177659 -0.3683433  1.3740320
##
## Residual standard error: 18.26047
## Degrees of freedom: 50 total; 48 residual
```

```
summary(abundance_lm_single)
```

```
## Generalized least squares fit by REML
##   Model: abundance_response ~ Cordia.lasiocalyx
##   Data: abundance_df
##        AIC      BIC    logLik
##   310.6455 318.1304 -151.3228
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~x + y
##  Parameter estimate(s):
##     range
## 78527.47
##
## Coefficients:
##                      Value Std.Error   t-value p-value
## (Intercept)       12.761349 129.59987 0.0984673   0.922
## Cordia.lasiocalyx  0.147346   0.20324 0.7249696   0.472
##
##  Correlation:
##                 (Intr)
## Cordia.lasiocalyx -0.018
##
## Standardized residuals:
##         Min         Q1         Med         Q3         Max
## -0.11056285 -0.10065151 -0.08725877 -0.05363536  0.19038239
##
## Residual standard error: 130.0813
## Degrees of freedom: 50 total; 48 residual
```

```
col_brks = hist(residuals(abundance_rat_nug_single), plot=F)$breaks
col_indices = as.numeric(cut(residuals(abundance_rat_nug_single), col_brks))
cols = rev(terrain.colors(length(col_brks)))
plot(BCI_xy, cex=2, pch=19, col=cols[col_indices])
```
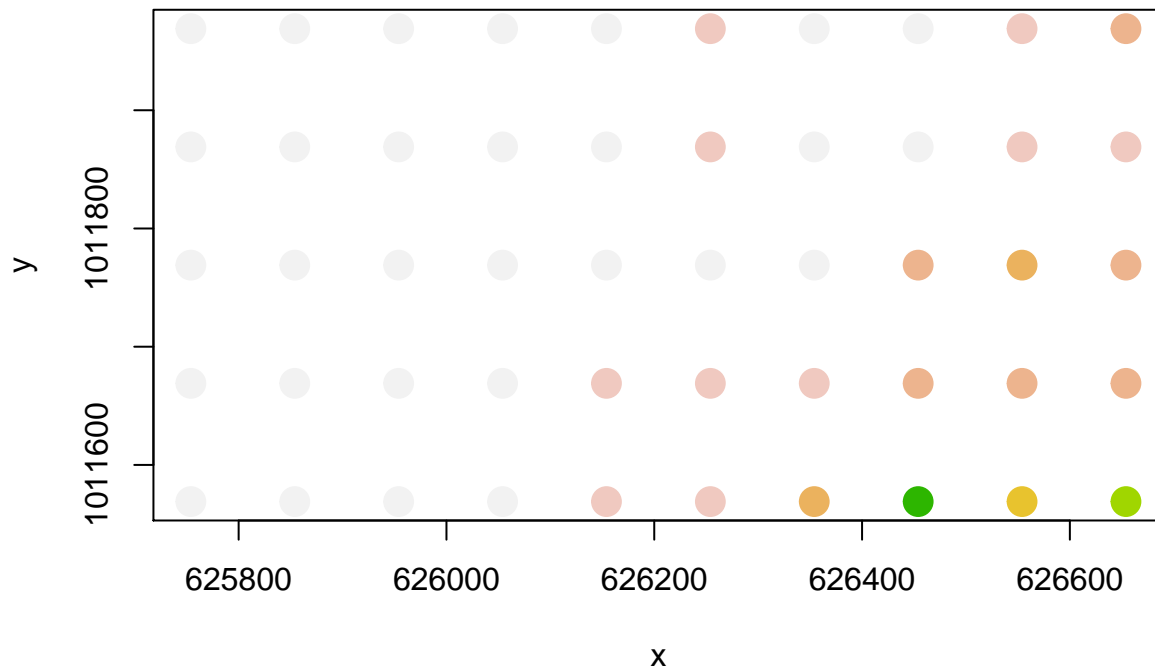
For Model 1, which includes Drypetes.standleyi as the predictor variable, the Variogram of the generalized linear model shows an upward trend which signifies spatial dependence of the residuals. In the Variogram of the normalized residuals there is somewhat of a trend present. The rational quadratic error model significantly improved model fit (supported by both its Variogram having the least variation and no trend and the anova results which show lowest AIC for this one of all 4 models).

Including the spatial error terms improved model fit as we can see from the anova() results. However, from the colored dots graph there does appear to still be spatial dependence.

Including the spatial error term did not have a large impact on the coefficients of the model, as seen from the summary() output of the rational quadratic model compared to that of the glm.

```
#model with all predictor variables- Model 2
formula1 = as.formula(paste("abundance_response ~ ", paste(sp_ids, collapse= "+")))

abundance_lm_all = gls(formula1, data=abundance_df,
        corr=corExp(form=~ x + y), control = glsControl(opt='optim',  msVerbose=T))

par(mfrow=c(1,1))
plot(Variogram(abundance_lm_all, form= ~ x + y))
```
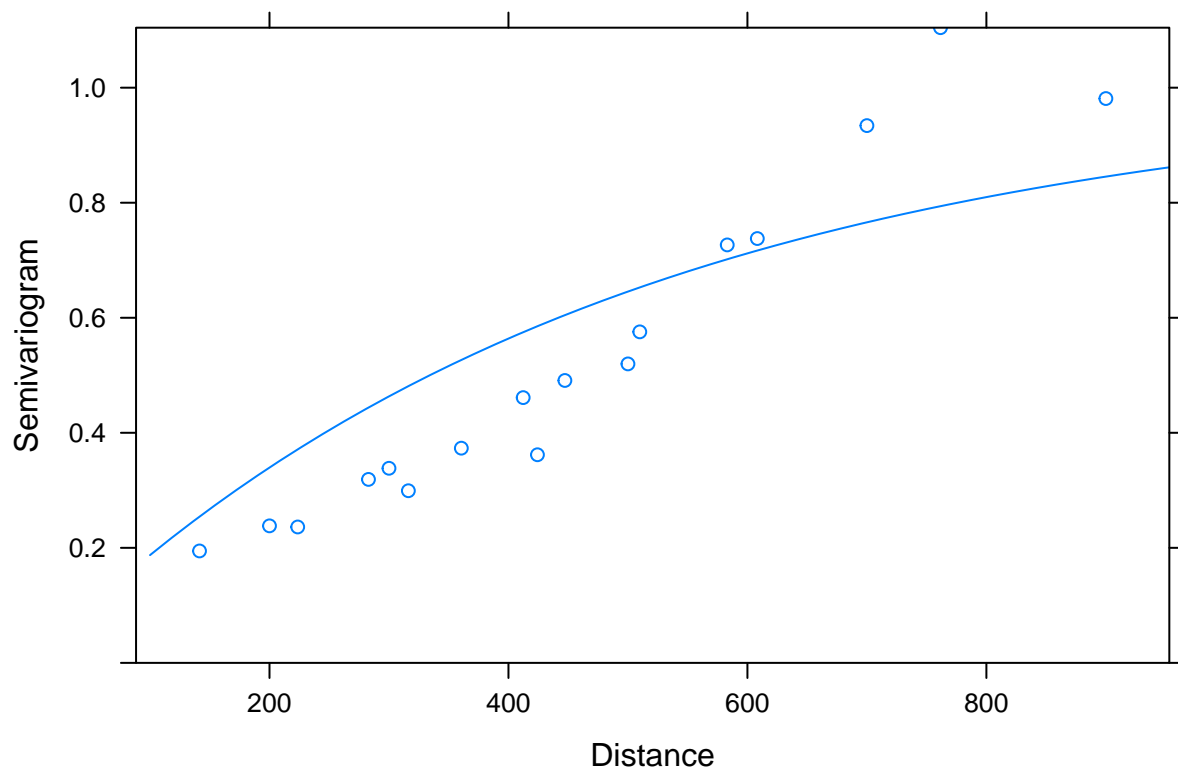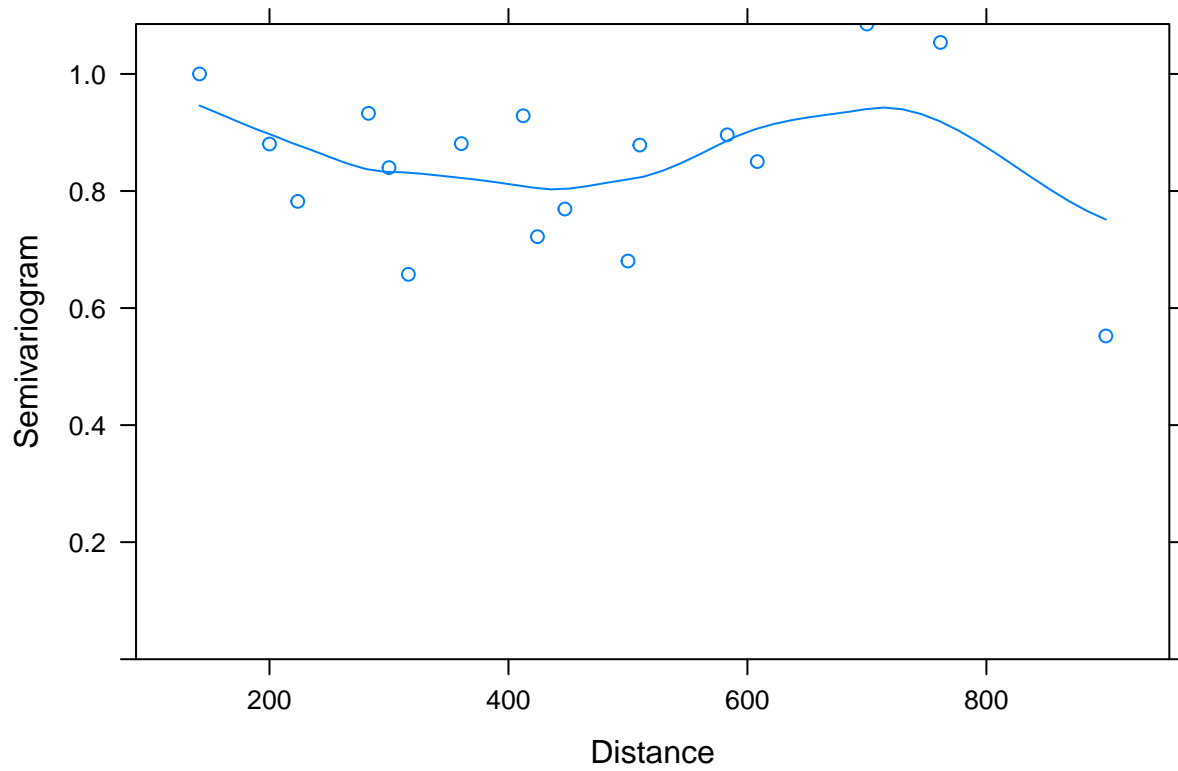
```
plot(Variogram(abundance_lm_all, resType='normalized', form= ~ x + y))
```

```r
#fit is good

abundance_exp_all = update(abundance_lm_all, corr=corExp(form=~x + y))
# examine fit of error model to the raw model residuals
#plot(Variogram(abundance_exp_all, maxDist = max_dist))
#worse fit
# examine the normalized residuals
#plot(Variogram(abundance_exp_all, resType='normalized', maxDist = max_dist))
#fit is good

# let's look at the same model but with a nugget
abundance_exp_nug_all = update(abundance_exp_all, corr=corExp(form=~x + y, nugget=T))
# Look to provide a better fit:
#plot(Variogram(abundance_exp_nug_all, maxDist = max_dist))
#plot(Variogram(abundance_exp_nug_all, resType = 'n', maxDist = max_dist))
#neither are very good

# let's examine the rational quadratic error model
abundance_rat_nug_all = update(abundance_lm_all, corr=corRatio(form=~x + y, nugget=T))
# examine fit of error model to model residuals
#plot(Variogram(abundance_rat_nug_all, maxDist = max_dist))
#plot(Variogram(abundance_rat_nug_all, resType='n', maxDist = max_dist))
#neither are very good

# let's compare the models
anova(abundance_lm_all, abundance_exp_all, abundance_exp_nug_all,
      abundance_rat_nug_all, test=F)
```

```
##                       Model df      AIC      BIC    logLik
## abundance_lm_all          1 10 301.6062 318.9829 -140.8031
## abundance_exp_all         2 10 301.6062 318.9829 -140.8031
## abundance_exp_nug_all     3 11 301.9683 321.0826 -139.9841
## abundance_rat_nug_all     4 11 303.1486 322.2630 -140.5743
```

```r
summary(abundance_lm_all)
```

```
## Generalized least squares fit by REML
##   Model: formula1
##   Data: abundance_df
##        AIC      BIC    logLik
##   301.6062 318.9829 -140.8031
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~x + y
##  Parameter estimate(s):
##     range
## 482.0443
##
## Coefficients:
##                      Value Std.Error  t-value p-value
## (Intercept)       2.3535050  6.171969 0.381322  0.7049
## Cordia.lasiocalyx 0.1207780  0.179794 0.671758  0.5054
```

```
## Hirtella.triandra        0.0190890   0.098504 0.193789   0.8473
## Picramnia.latifolia      0.2011459   0.509152 0.395061   0.6948
## Quassia.amara            1.2771778   1.847478 0.691309   0.4932
## Tabernaemontana.arborea  0.0675475   0.133767 0.504963   0.6162
## Trattinnickia.aspera     1.8116442   0.525080 3.450226   0.0013
## Xylopia.macrantha        0.3386982   0.156876 2.159022   0.0366
##
##   Correlation:
##                          (Intr) Crd.ls Hrtll. Pcrmn. Qss.mr Tbrnm. Trttn.
## Cordia.lasiocalyx        -0.225
## Hirtella.triandra        -0.308 -0.022
## Picramnia.latifolia       0.045 -0.066 -0.369
## Quassia.amara            -0.059 -0.303  0.322 -0.142
## Tabernaemontana.arborea  -0.239 -0.016  0.288 -0.221  0.112
## Trattinnickia.aspera     -0.069  0.168 -0.237  0.212 -0.633 -0.041
## Xylopia.macrantha        -0.056 -0.137 -0.063  0.109  0.290  0.102 -0.185
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -1.0036421 -0.5230822 -0.3173885  0.2201812  2.3702271
##
## Residual standard error: 8.644825
## Degrees of freedom: 50 total; 42 residual
```

```r
summary(abundance_exp_nug_all)
```

```
## Generalized least squares fit by REML
##   Model: formula1
##   Data: abundance_df
##        AIC      BIC    logLik
##   301.9683 321.0826 -139.9841
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~x + y
##  Parameter estimate(s):
##        range       nugget
## 2.987795e+04 4.384772e-03
##
## Coefficients:
##                              Value Std.Error    t-value p-value
## (Intercept)              3.0176507  41.75316  0.072274  0.9427
## Cordia.lasiocalyx        0.1426934   0.18954  0.752855  0.4557
## Hirtella.triandra       -0.0014346   0.09043 -0.015863  0.9874
## Picramnia.latifolia      0.2863453   0.52745  0.542883  0.5901
## Quassia.amara            1.3311569   1.93645  0.687420  0.4956
## Tabernaemontana.arborea  0.0407405   0.13953  0.291994  0.7717
## Trattinnickia.aspera     1.8165108   0.57285  3.171032  0.0028
## Xylopia.macrantha        0.4085102   0.15374  2.657119  0.0111
##
##   Correlation:
##                          (Intr) Crd.ls Hrtll. Pcrmn. Qss.mr Tbrnm. Trttn.
## Cordia.lasiocalyx        -0.040
## Hirtella.triandra        -0.044 -0.098
## Picramnia.latifolia       0.004  0.016 -0.360
```

```
## Quassia.amara            -0.008 -0.292  0.344 -0.193
## Tabernaemontana.arborea -0.031 -0.020  0.160 -0.197  0.088
## Trattinnickia.aspera     -0.014  0.165 -0.276  0.255 -0.655 -0.036
## Xylopia.macrantha        -0.008 -0.066 -0.038 -0.047  0.306  0.140 -0.183
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -0.20840996 -0.12258056 -0.08687393  0.02426672  0.44015110
##
## Residual standard error: 42.12228
## Degrees of freedom: 50 total; 42 residual
```

```
col_brks = hist(residuals(abundance_lm_all), plot=F)$breaks
col_indices = as.numeric(cut(residuals(abundance_lm_all), col_brks))
cols = rev(terrain.colors(length(col_brks)))
plot(BCI_xy, cex=2, pch=19, col=cols[col_indices])
```
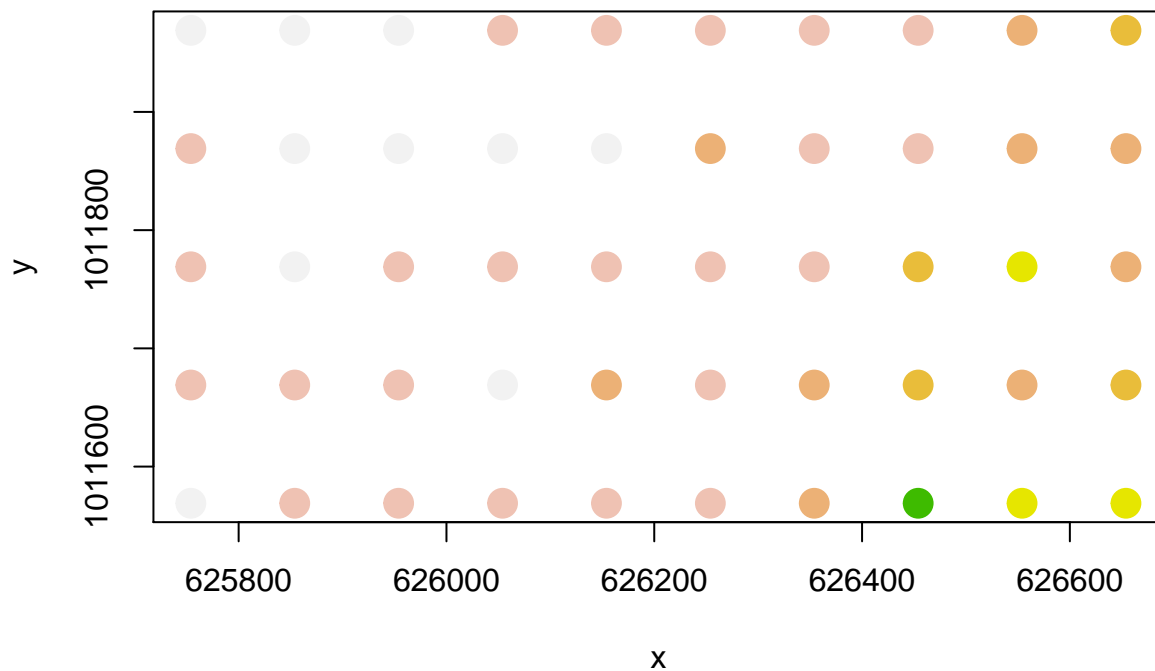

For Model 2, which includes multiple species as predictor variables, the Variogram of the generalized linear model shows an upward trend which signifies spatial dependence of the residuals, but the line has very good fit, and after normalizing the residuals the residuals appear mostly normally distributed with little trend. The generalized linear model performs the best/equally as good as the exponential model over the ones with nuggets (supported by both its Variogram having the least variation and no trend and the anova results which show lowest AIC for this model).

Improving the spatial error terms did not improve model fit as we can see from the anova() results, since the generalized linear model performed just as well if not better than the others. From the colored dots graph there does not appear to be a pattern of spatial dependence with this model. It also did not have a large impact on the coefficients of the model, which can be seen when comparing the summary() output of the glm model vs. the exponential model with a nugget, for example.

Modeling the spatial errors has a large influence in the first model but not in the second model because the second model has many more predictor variables, so their individual effects/spatial relationships with the response variable (if some have spatial dependence) are less influential because of the presence of many

other predictor variables. However, in the case of model 1 where a spatially dependent variable was the only predictor in the model, it had a large influence on the model.