Carbon Dioxide Predictions by Location

Hanna Grossman

3/18/2020

1. Introduction

Through this project, I will explore one of the National Oceanic and Atmospheric Administration's data sets, focusing in on carbon cycle research. This dataset contains data measurements collected from the Barnett Shale region of Texas. These variables include latitude, longitude, altitude, and height that each measurment was taken. The measurements include the mole fraction of CH4, CO, CO2, and Ethane in dry air. The air pressure and temperature were also recorded for each observation. I will begin with an exploratory analysis of each variable mentioned above. This will include summary statistics, histograms, scatterplots, variance covariance matrices, correlation matrices, and bubble plots. I will also compute sample variograms for each variable and fit theoretical variograms to them. In addition, I will compute cross-semivariograms for pairs of variables. From there, I will next focus in on the target variable, or variable interest, carbon dioxide (CO2). I chose this variable as the target variable, because this data was collected for the purpose of researching the carbon cycle, and I therefore believe predicting the amount of CO2 based on location, and other co-located variables, may prove quite interesting and directly align with the purpose of this data collection. With this variable, I will cross validate to find which model variogram is the best fit. Through this process, I find that the spherical model is the best fit for the CO2 variable, as it results in a lower PRESS when compared to the exponential and linear models during cross validation. From there, I perform ordinary kriging, universal kriging, and co-kriging in order to predict our CO2 variable. For co-kriging, I use carbon monoxide (CO), as the co-located variable. After performing these three types of kriging, I then use cross validation to compare the three methods to each other. I find that co-kriging performs best, and therefore this will be my chosen method of kriging to allow for predicting the CO2 variable. From there, I construct both a raster map of the predicted values, and a raster map of the kriging variances, both with contours, for my chosen method, co-kriging. In the end, I am able to gain a strong understanding of the data as a whole through the exploratory data analysis and variograms. I successfully predict carbon dioxide values in the Barnett Shale region of Texas, using an spherical model to fit the sample variogram, and co-kriging to compute my predictions. Finally, through examining the raster map of predicted values, I observe that the lowest carbon dioxide levels are found in the southwest area of the Barnett Shale region of Texas, while the highest carbon dioxide levels are found in the northeast and central east areas of the Texas region. I also observe through the raster map of variances that there is greater variance in some of the darker orange areas, for example around longitude -97.5 and latitude 32.5.

2. Data

Data source

The data used in this study can be found through the National Oceanic and Atmospheric Administration (NOAA) website, through the Earth System Research Laboratory Global Monitoring Division (ESRL/GMD) Data Finder. The link to the data is cited below, and the data will also be submitted along with this report.

https://www.esrl.noaa.gov/gmd/dv/data/index.php?search=coordinate

Describing data

This dataset consists of airborne measurements, taken in the Barnett Shale region of Texas. These measurements were taken to allow for carbon-cycle research. The measurements include the longitude and latitude (renamed x and y in the data cleaning step) that the measurements were taken, the altitude and intake height of each measurement, the mole fraction of CH4, CO, CO2, and Ethane in dry air, and the air pressure and air temperature at each measurement location.

Exploratory data analysis

Performing non-spatial analysis of data

Descriptive statistics:

```
summary(Barnett_sub$x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -100.02  -98.20  -97.58  -97.58  -96.90  -95.51
```

```
summary(Barnett_sub$y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31.88   32.23   33.18   33.08   33.75   34.84
```

```
summary(Barnett_sub$altitude)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   197.3   684.3   779.1   903.0  1059.0  2983.4
```

```
summary(Barnett_sub$intake_height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.21  408.38  539.44  642.25  752.32 2682.15
```

```
summary(Barnett_sub$CH4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1842    1936    1957    1962    1984    2374
```

```
summary(Barnett_sub$CO)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   74.71  120.86  134.25  135.69  147.55  227.51
```

```
summary(Barnett_sub$CO2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   392.6   397.4   399.8   400.0   403.0   416.2
```

```
summary(Barnett_sub$Ethane)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.030   3.995   5.705   5.721   6.963  21.800
```

```
summary(Barnett_sub$P)
```

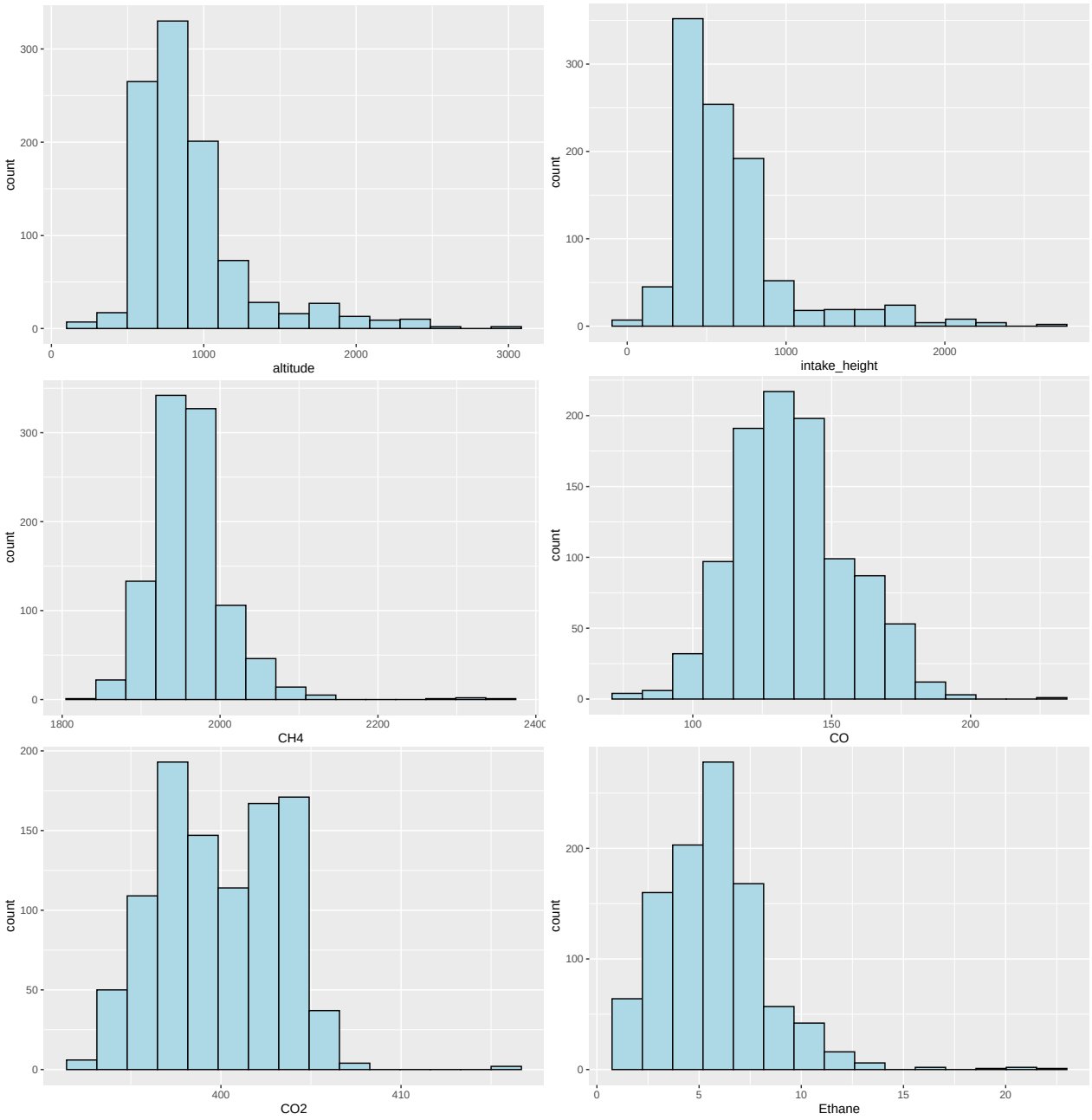
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   71639   90171   92738   91757   94380   99497
```

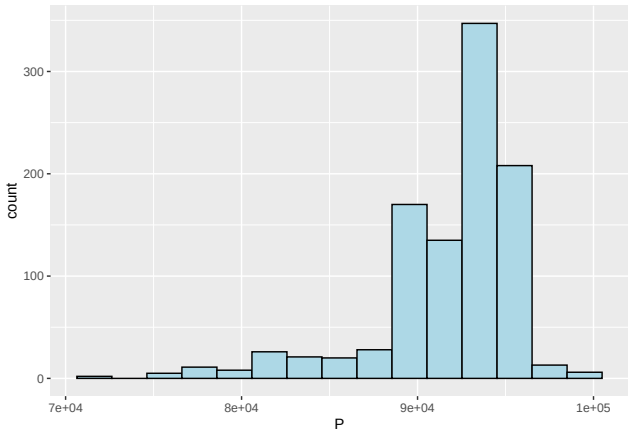
```
summary(Barnett_sub$T)
```

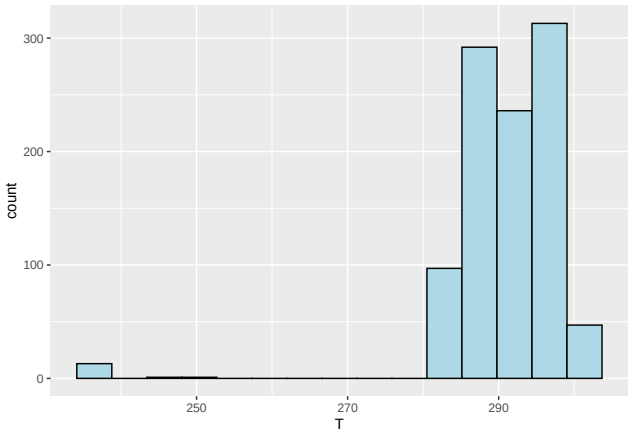
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   237.2   287.3   291.5   290.9   295.8   302.1
```

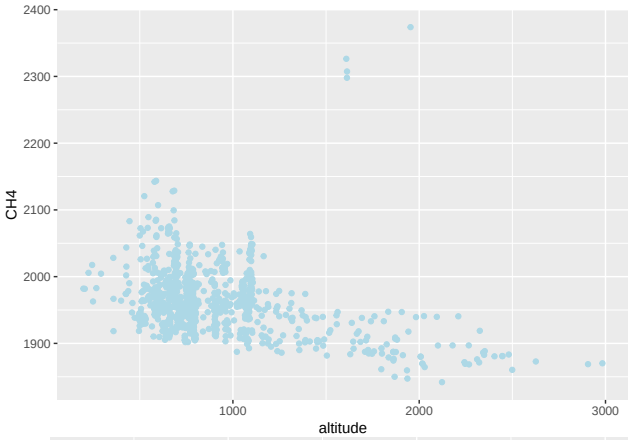
Histograms

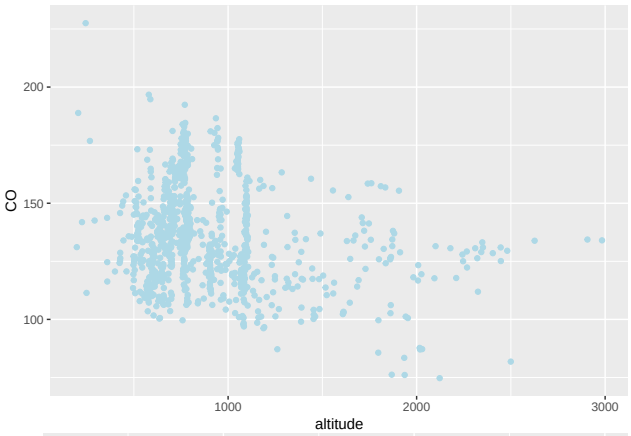


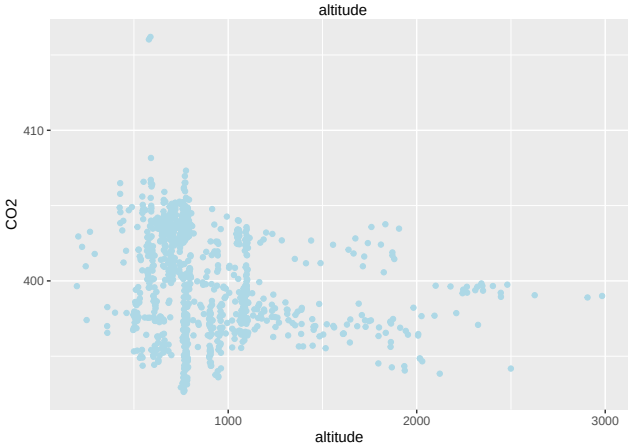


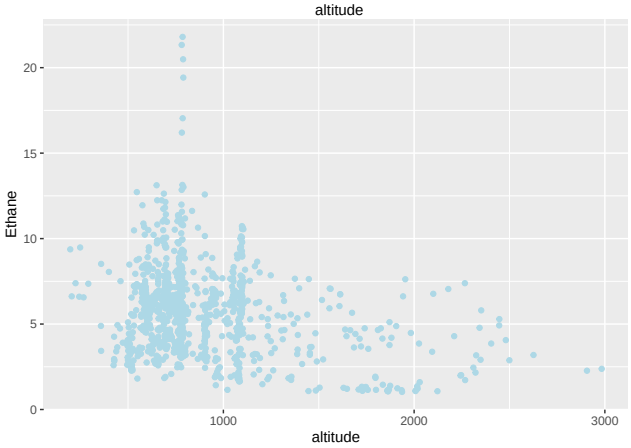


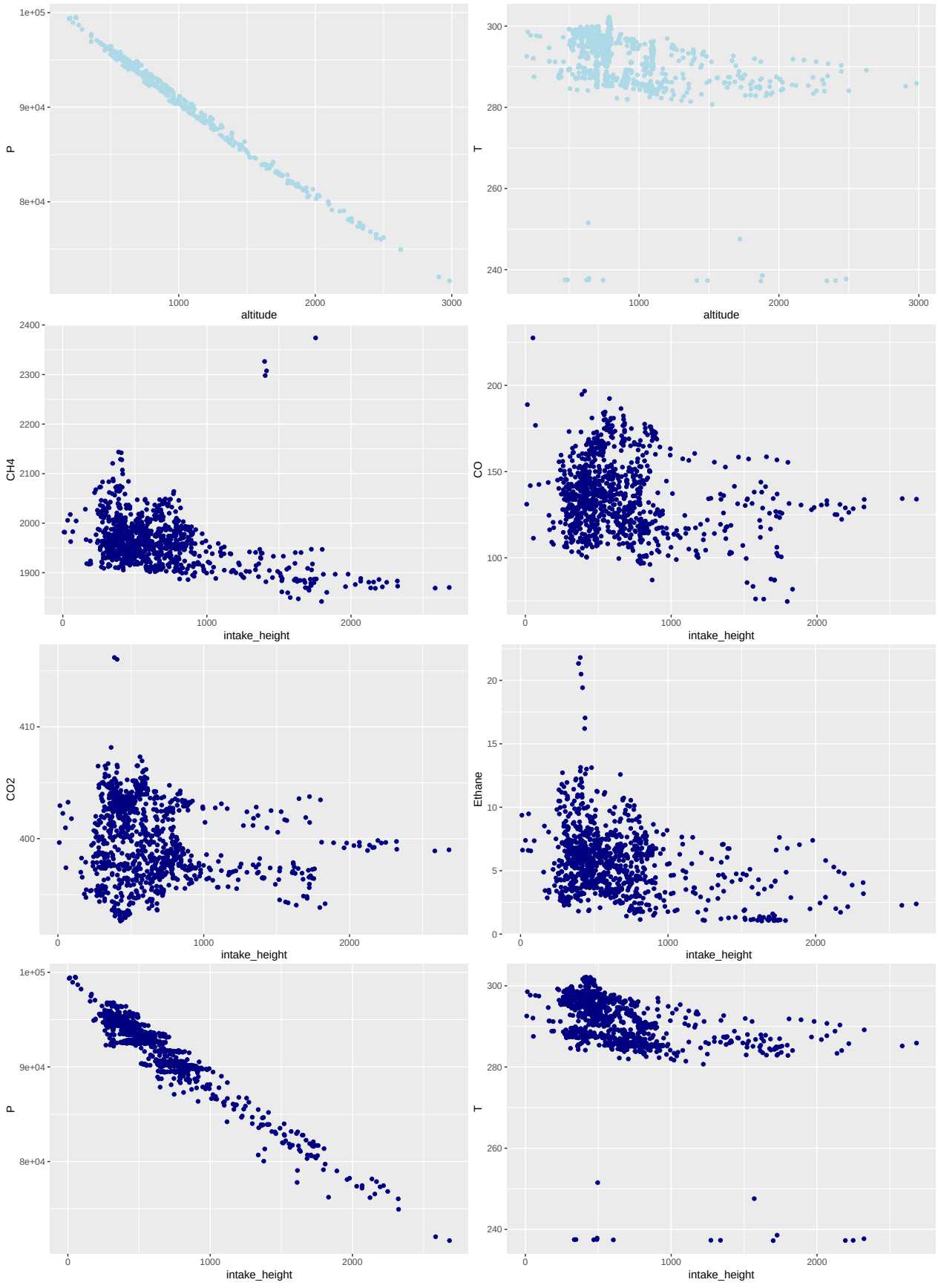
Scatterplots










P
altitude
T
altitude
CH4
intake_height
CO
intake_height
CO2
intake_height
Ethane
intake_height
P
intake_height
T
intake_height

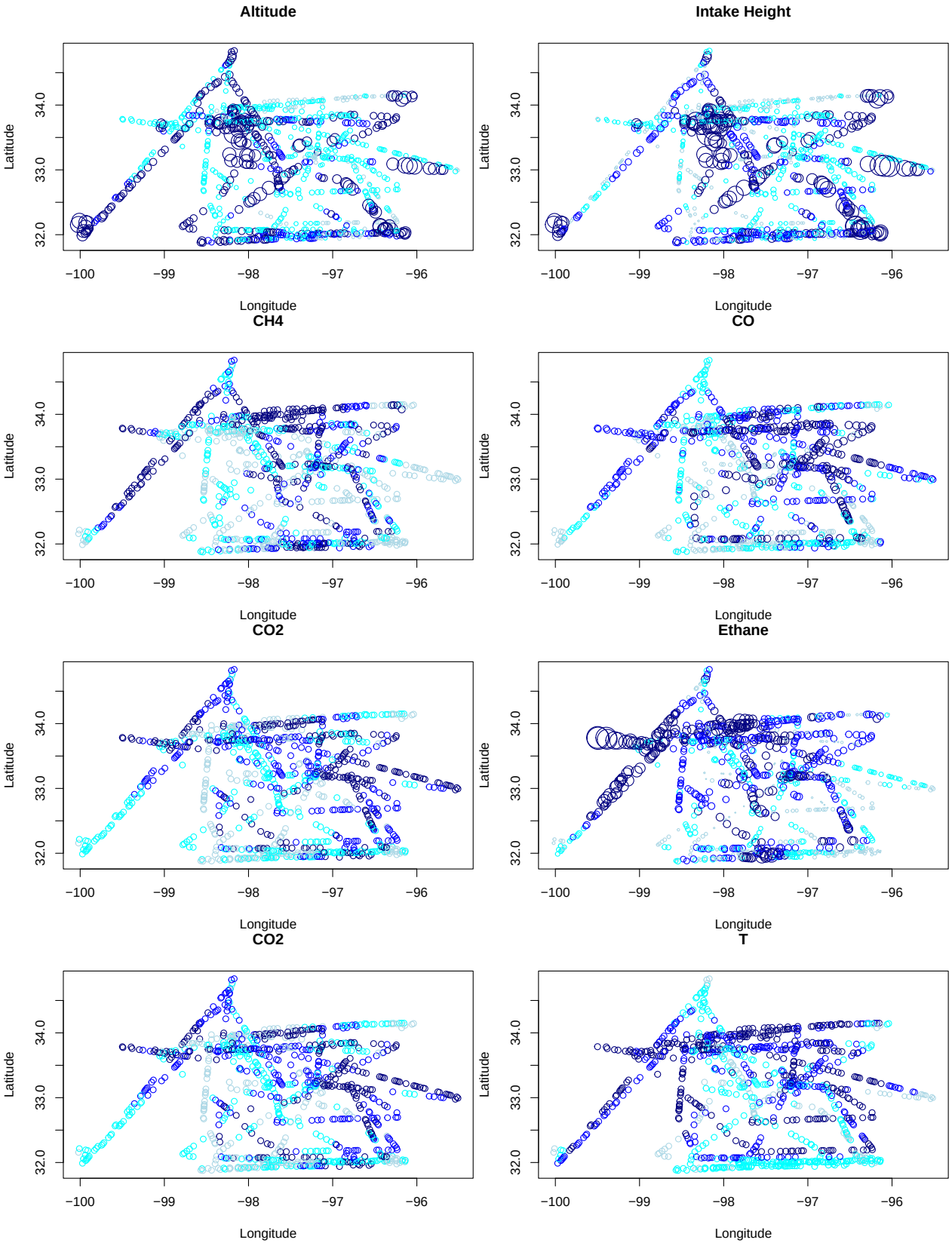
Variance covariance matrix

```
##                    altitude intake_height         CH4          CO         CO2
## altitude        146448.3493   139254.4853 -6363.42053 -1811.43299 -312.329228
## intake_height   139254.4853   142161.3585 -6604.14518 -1505.39293 -243.613234
## CH4              -6363.4205    -6604.1452  2434.58880   361.67977   51.553732
## CO               -1811.4330    -1505.3929   361.67977   415.81445   42.983284
## CO2                -312.3292     -243.6132    51.55373    42.98328   11.387065
## Ethane             -269.3349     -327.1250    73.27401    16.39076    2.553961
## P             -1511505.1987 -1443102.3858 68648.36777 20075.49857 3738.620382
## T                 -1133.0392    -1276.8606    66.44042    22.26751    1.623082
##                     Ethane            P            T
## altitude      -269.334934 -1511505.199 -1133.039150
## intake_height -327.124979 -1443102.386 -1276.860565
## CH4             73.274012    68648.368    66.440424
## CO              16.390765    20075.499    22.267513
## CO2              2.553961     3738.620     1.623082
## Ethane           6.590465     2911.422     5.998183
## P             2911.421699 15745983.051 11423.691952
## T                5.998183    11423.692    64.670077
```

Correlation matrix

```
##                  altitude intake_height        CH4         CO         CO2
## altitude        1.0000000     0.9651086 -0.3370044 -0.2321295 -0.24186024
## intake_height   0.9651086     1.0000000 -0.3549875 -0.1957985 -0.19147152
## CH4            -0.3370044    -0.3549875  1.0000000  0.3594692  0.30962883
## CO             -0.2321295    -0.1957985  0.3594692  1.0000000  0.62466029
## CO2            -0.2418602    -0.1914715  0.3096288  0.6246603  1.00000000
## Ethane         -0.2741526    -0.3379597  0.5784677  0.3131060  0.29481545
## P              -0.9953653    -0.9645427  0.3506166  0.2481028  0.27920332
## T              -0.3681722    -0.4211153  0.1674433  0.1357908  0.05981125
##                   Ethane          P           T
## altitude      -0.2741526 -0.9953653 -0.36817219
## intake_height -0.3379597 -0.9645427 -0.42111528
## CH4            0.5784677  0.3506166  0.16744335
## CO             0.3131060  0.2481028  0.13579076
## CO2            0.2948155  0.2792033  0.05981125
## Ethane         1.0000000  0.2858000  0.29054272
## P              0.2858000  1.0000000  0.35798918
## T              0.2905427  0.3579892  1.00000000
```

Creating circle (bubble) plots to show the data against the coordinates



3. Methodology

The methodology is listed in steps below, along with an explanation of each step. The results for steps 2 and 3 can be found above in the data section, and the results for steps 4-10 can be found below in the results section. These results consist of both output and plots.

1. Reading in, cleaning, and subsetting data
   After reading in the data, and renaming the columns to their correct names, I then filter the data to keep only unique locations. I use the distinct function to keep only one observation for each location. From there, I remove all negative values for CH4, CO, CO2, and Ethane. Because these variables are measurements of how much each of these chemical compunds are found in a given location, they cannot be negative, and negative values therefore are values that were inputted incorrectly. Finally, I take a random sample of 1000 observations to end up with a manageable dataset, and keep only the columns I will use throughout my analysis.

2. Non-spatial analysis of data
   I next perform a non-spatial analysis of the data in order to better understand what the data look like, and to see how this may impact the rest of my analysis. This exploratory analysis consists of descripive statistics using the summary function, histograms of each variable. I also plot scatterplots with altitude and intake height as the x variables, and CH4, CO, CO2, and Ethane, the variables measured in the data set, as the y variables. This allows me to see if the amounts of these chemical compounds are affected by the altitude and height of the location. Finally, I compute the variance covariance matrix and the correlation matrix to better understand how the variables are related to one another.

3. Creating circle (bubble) plots to show the data against the coordinates
   In this step, I create plots with the x axis as longitude, and the y axis as latitude. I make each point a color based on where its value is compared to each variables' (altitude, intake height, CH4, CO, CO2, ethane, air pressure, air temperature) min, Q1, median, Q3, and max. I also change the size of each point based on how the value of each point compares to the mean of that variable. These plots can be found above in the data section as well.

4. Computing semivariograms for all variables
   In this step, I compute semivariograms for each variable. These variables include: altitutde, intake height, methane (CH4), carbon monoxide (CO), carbon dioxide (CO2), ethane, air pressure (P), and air temperature (T). In order to compute the variograms, I use the package gstat, with functions gstat, variogram, fit.variogram, and plot. To find the best fit, I try transforming variables, for example using a log transformation, using different directions for the variograms, and removing trends. I also try fitting different theoretical models to each variogram, including spherical, gaussian, exponential, and linear. You can see how I end up plotting each variogram specifically in the code show below in the results section.

5. Computing cross-semivariograms for pairs of variables
   To compute cross-semivariograms for pairs of variables, I once again use the gstat packages with functions gstat, variogram, fit.lmc, and plot. I plot variables that seem to compute the best cross-semivariograms together. This includes altitude with intake height, CO with CO2, CH4 with ethane, and air pressure with air temperature.

6. Choosing a target variable to focus in on
   After completing the exploratory data analysis and computing the variograms for each variable, along with gaining an understanding for the dataset, I chose a target variable to focus in on, CO2. Since this data was collected for the purpose of researching the carbon cycle, I am choosing to focus in on the variable Carbon Dioxide (CO2). Predicting the amount of CO2 based on location may prove to be quite interesting, and will directly align with the purpose of this data collection. I will see how carbon dioxide varies depending on location, and if this allows me to predict the variable well based on a given location. In addition, I will be able to explore how other variables, for example carbon monoxide, improve my predictions of carbon dioxide.

7. Performing cross validation to choose best model variogram

- Dividing the data set into two parts - one for modeling and one for cross validations
  I begin by dividing the dataset into two, a training and a validation set. The training set contains 70% of the data, or 700 of the 1,000 observations, and the validation set contains the other 30% of the data. From there, I plot the sample variogram using the training dataset, and then fit the spherical, exponential, and linear models to the sample variogram. Next, I use the krige function to find the predicted values for CO2 using each of the three models. Finally, I compute the PRESS for each model, by calculating the sum of squared differences between the true CO2 values and the predicted CO2 values. The best fit model is the one with the lowest PRESS. Therefore, through cross validation I am able to find which model is the best fitting and proceed using this model type.

8. Performing ordinary, universal, and co-kriging, using CO as the co-located variable for co-kriging
   I perform ordinary and universal kriging using the krige function. For ordinary kriging, I specify that formula=CO2~1, while for universal kriging, I specify that formula=CO2~x+y. I am then able to view the predictions made through each form of kriging looking at CO2.pred, produced by the krige function. To perform co-kriging, I use the predict function, with the vm.fit and grid objects I create previously. I once again can view the predictions by looking at CO2.pred, produced by the predict function.

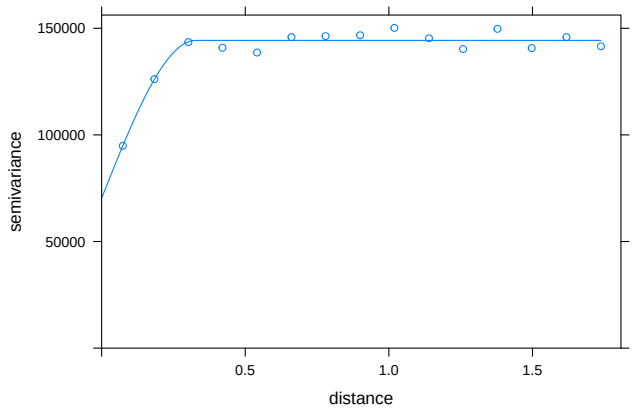
9. Performing cross validation to choose which type of kriging performs best
   After performing ordinary, universal, and co-kriging, I perform cross validation to find which type of kriging is best. I use krige.cv to find the sum of squared residuals for ordinal and universal kriging, and I use gstat.cv to find the sum of squared residuals for co-kriging. The method with the lowest sum of squared residuals therefore performs the best, and is the method of kriging I will choose to move forward with, using the CO2 predictions found through this method.

10. Constructing a raster map of the predicted values and a raster map of the kriging variances, adding contours to these maps, using the method of kriging that performs best above
    Finally, I construct raster maps of the predicted values and of the kriging variances, first by collapsing the predicted values, or the variances, into a matrix. I then use the image function to create the initial image, and then use the contour function to add in the contours. Finally, for the raster map of the predicted values, I graph the points using the points function. I do this all for co-kriging, with the spherical model, as we found these perform best through cross validation.

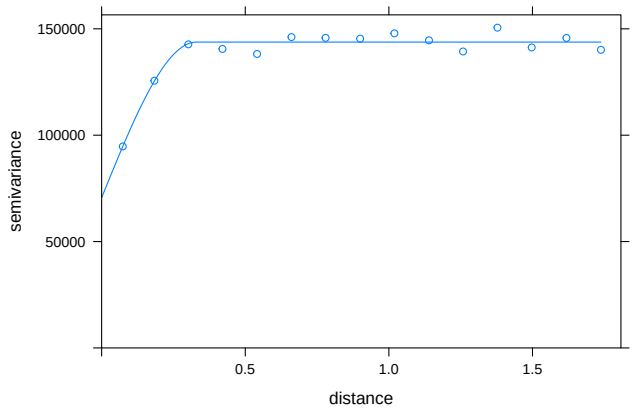
4. Results

Computing semivariograms for all variables

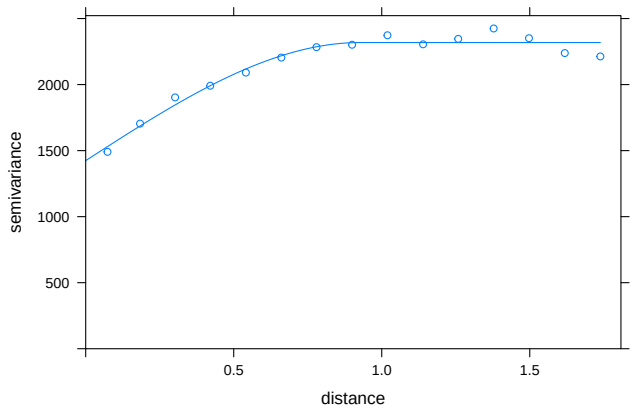
```
#altitude
g1 <- gstat(id="altitude", formula=altitude~1, locations=~x+y,
            data=Barnett_sub)
vario1 <- variogram(g1)
v.fit1 <- fit.variogram(vario1, vgm(psill=100000, model="Sph",
                                    range=0.5, nugget=75000), fit.method=6)
plot(vario1, v.fit1)
```



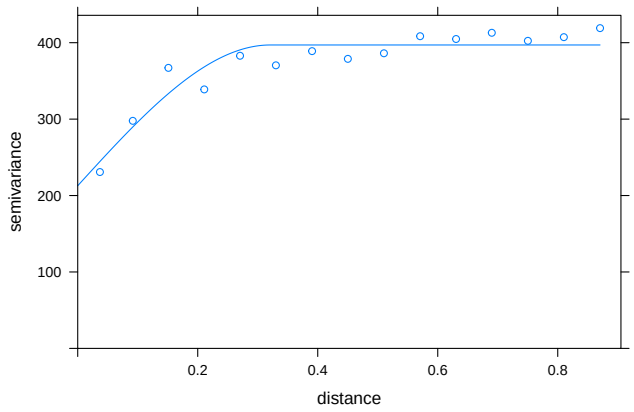

```
#intake height
g2 <- gstat(id="intake_height", formula=intake_height~1,
            locations=~x+y, data=Barnett_sub)
vario2 <- variogram(g2)
v.fit2 <- fit.variogram(vario2, vgm(psill=100000, model="Sph",
                                    range=0.5, nugget=75000), fit.method=6)
plot(vario2, v.fit2)
```



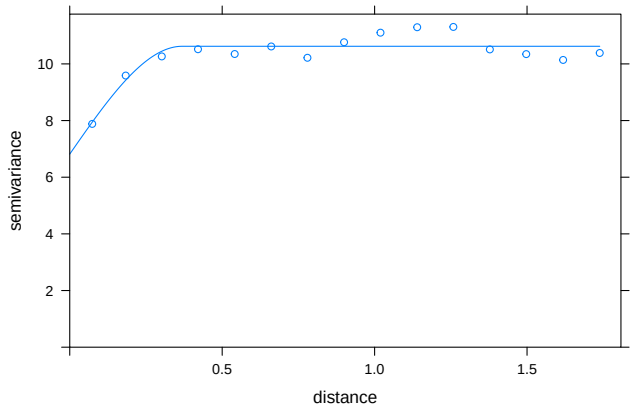

```
#CH4
g3 <- gstat(id="CH4", formula=CH4~1,
            locations=~x+y, data=Barnett_sub)
vario3 <- variogram(g3)
v.fit3 <- fit.variogram(vario3, vgm(psill=1250, model="Sph",
                                    range=1, nugget=1500), fit.method=6)
plot(vario3, v.fit3)
```



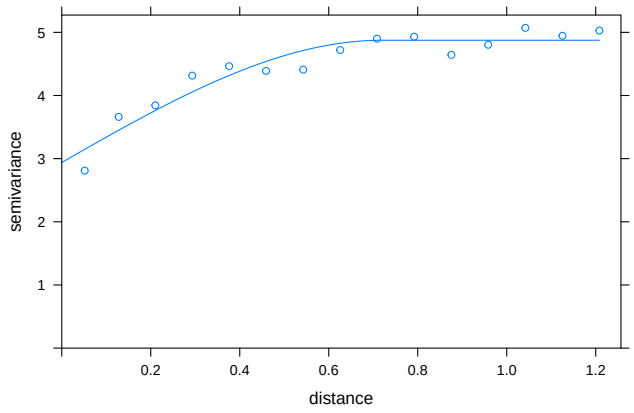
```
#CO
g4 <- gstat(id="CO", formula=CO~1,
            locations=~x+y, data=Barnett_sub)
vario4 <- variogram(g4, cutoff=0.9)
v.fit4 <- fit.variogram(vario4, vgm(psill=300, model="Sph",
                                    range=0.5, nugget=100), fit.method=6)
plot(vario4, v.fit4)
```



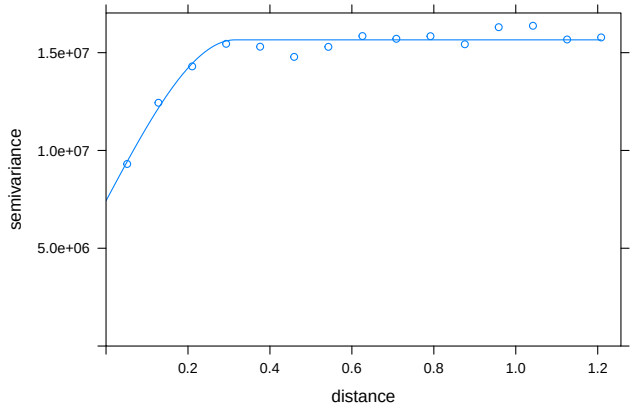

```
#CO2
g5 <- gstat(id="CO2", formula=CO2~x+y,
            locations=~x+y, data=Barnett_sub)
vario5 <- variogram(g5)
v.fit5 <- fit.variogram(vario5, vgm(psill=4.5, model="Sph",
                                    range=.5, nugget=7.5), fit.method=6)
plot(vario5, v.fit5)
```



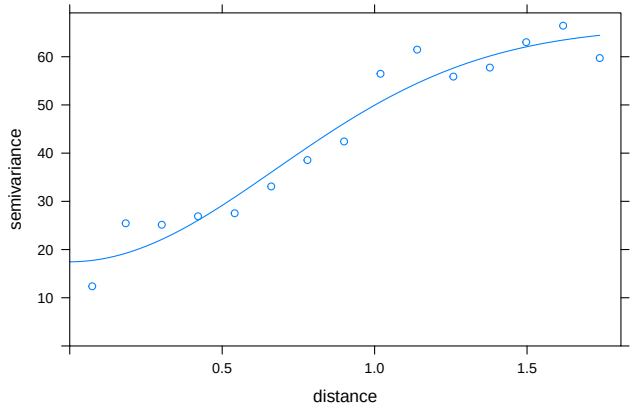

```
#Ethane
g6 <- gstat(id="Ethane", formula=Ethane~x+y, locations=~x+y, data=Barnett_sub)
vario6 <- variogram(g6, cutoff=1.25)
v.fit6 <- fit.variogram(vario6, vgm(psill=3.2, model="Sph",
                                    range=1, nugget=2), fit.method=6)
plot(vario6, v.fit6)
```



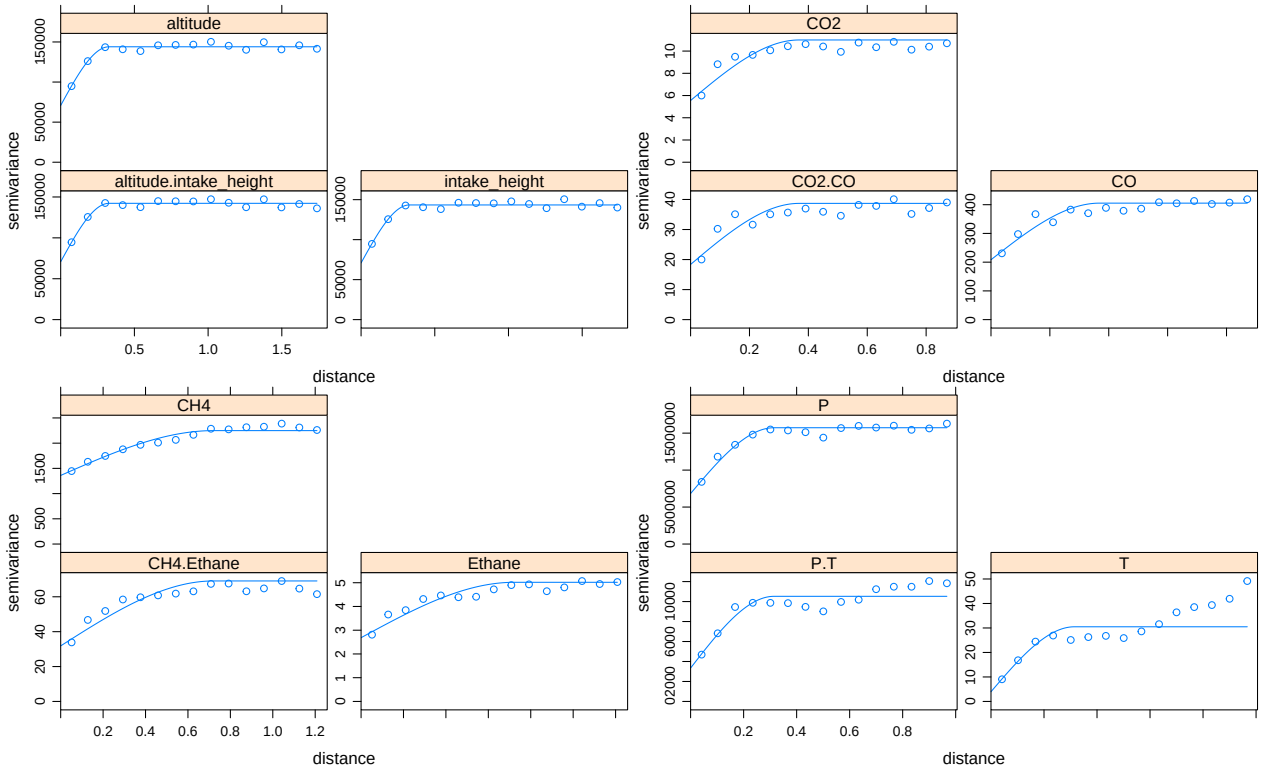

```
#P
g7 <- gstat(id="P", formula=P~1, locations=~x+y, data=Barnett_sub)
vario7 <- variogram(g7, cutoff=1.25)
v.fit7 <- fit.variogram(vario7, vgm(psill=1.5e+07, model="Sph",
                                    range=0.75, nugget=5e+06), fit.method=6)
plot(vario7, v.fit7)
```




```
#T
g8 <- gstat(id="T", formula=T~1, locations=~x+y, data=Barnett_sub)
vario8 <- variogram(g8)
v.fit8 <- fit.variogram(vario8, vgm(psill=900, model="Gau",
                                    range=1, nugget=400), fit.method=6)
plot(vario8, v.fit8)
```



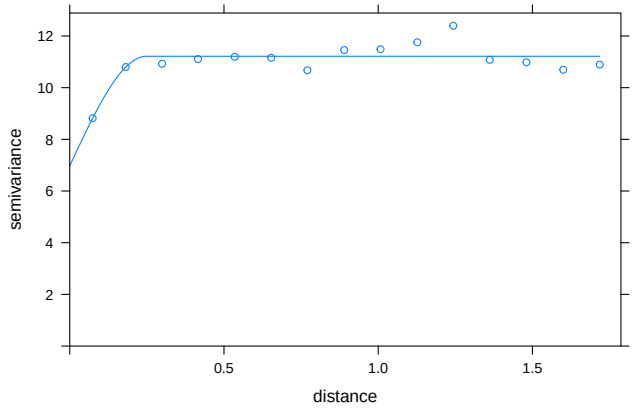
Computing cross-semivariograms for each pair of variables



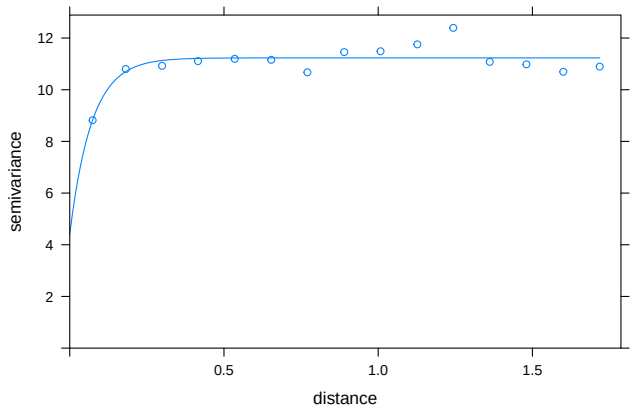
Performing cross validation to choose between different types of model variograms

Cross validation: dividing the data set into two parts - one for modeling and one for cross validations

```
#spherical variogram
v.fit_sph <- fit.variogram(vario, vgm(psill=4.5, model="Sph",
                                      range=.5, nugget=7.5), fit.method=6)
plot(vario, v.fit_sph)
```



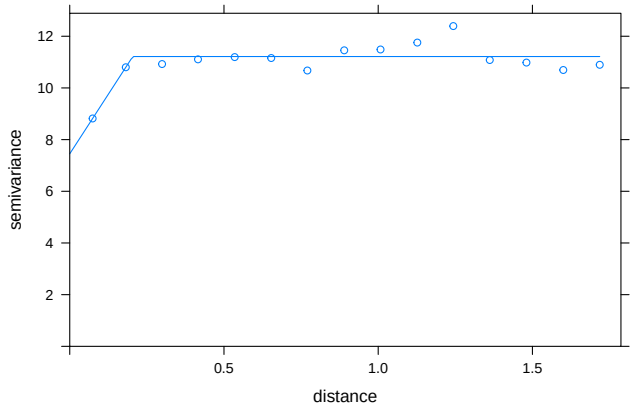

```
#exponential variogram
v.fit_exp <- fit.variogram(vario, vgm(psill=4.5, model="Exp",
                                      range=.5, nugget=7.5), fit.method=6)
plot(vario, v.fit_exp)
```




```
#linear variogram
v.fit_lin <- fit.variogram(vario, vgm(psill=4.5, model="Lin",
                                      range=.5, nugget=7.5), fit.method=6)
```

```
## Warning in fit.variogram(vario, vgm(psill = 4.5, model = "Lin", range = 0.5, :
## linear model has singular covariance matrix
```

```
plot(vario, v.fit_lin)
```

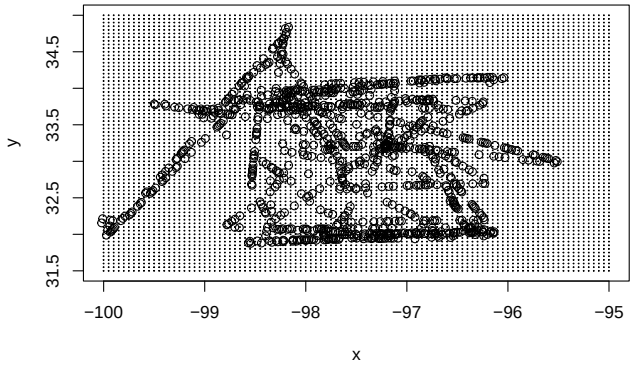



```
## [using ordinary kriging]
## [1] 400.9145 398.8284 401.3153 399.9590 398.6181 399.2799
## [using ordinary kriging]
## [1] 401.1419 398.8825 401.3507 399.8060 399.1238 399.1095
## [using ordinary kriging]
## [1] 400.9427 398.6879 401.5328 399.8431 398.6025 399.3182
## PRESS for spherical model: 1772.32168151433
## PRESS for exponential model: 1779.84608211147
## PRESS for linear model: 1789.23361006402
```

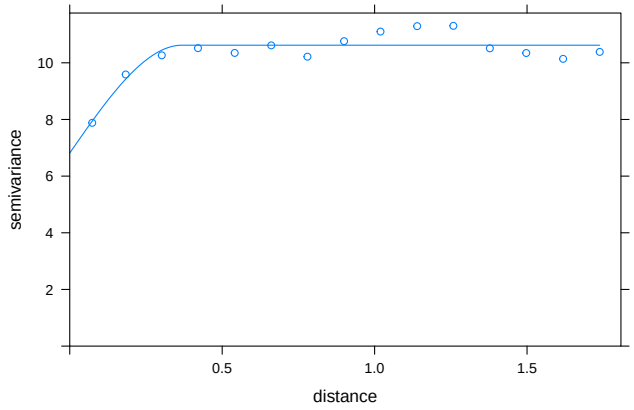
After performing cross validation, I choose the spherical model, as it results in a lower PRESS when compared to the exponential and linear models.

Setting Up for Krigging

Constructing the grid for kriging predictions



Fitting the spherical variogram for the CO2 variable



Performing ordinary kriging

```
## [using ordinary kriging]
## head of predictions for ordinary kriging:
## [1] 399.9509 399.9509 399.9509 399.9509 399.9509 399.9509
```

Performing universal kriging

```
## [using universal kriging]
## head of predictions for universal kriging:
## [1] 396.5917 396.6323 396.6729 396.7136 396.7542 396.7948
```

Performing co-kriging

```
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## head of predictions for co-kriging:
## [1] 398.7312 398.7532 398.7753 398.7973 398.8194 398.8414
```

Performing cross validation to choose between types of kriging

```
## the sum of squared residuals for ordinary kriging: 6604.09600004152
```

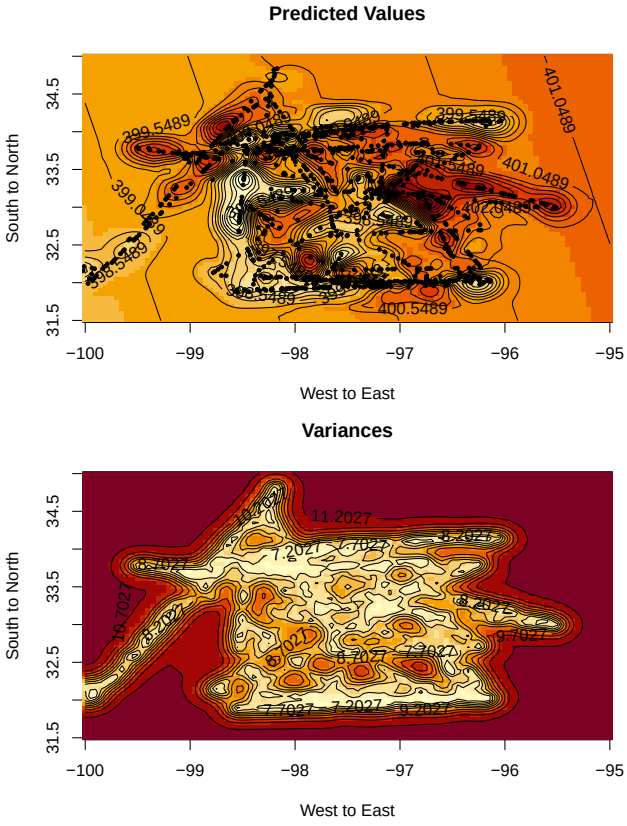
```
## the sum of squared residuals for universal kriging: 6584.37566635749
```

```
## the sum of squared residuals for co-kriging: 4461.58334900544
```

From the cross validation above, we see that co-kriging performs best by far. The next best is universal kriging, and finally, ordinary kriging.

Constructing a raster map of the predicted values and a raster map of the kriging variances, adding contours to these maps

I construct the raster maps using co-kriging, as we see above through cross validation that this method performs best.



5. Conclusion

Through the raster map above of predicted values, we see that the lowest carbon dioxide levels are found in the southwest area of the Barnett Shale region of Texas. Then moving in a diagonal direction, we see that the highest carbon dioxide levels are found in the northeast and central east areas of the Texas region. It is important to note that we should only be predicting carbon dioxide levels within the scope of our data, as kriging is an interpolator and not an extrapolator. We see through our raster map of variances that there is greater variance in some of the darker orange areas, for example around longitude -97.5 and latitude 32.5. Of course, the variance is high in the red areas, as there are no data there.
In conclusion, through this project, I am able to gain a strong understanding of the data as a whole through the exploratory data analysis and variograms. I am then able to predict carbon dioxide levels through co-kriging, using the latitude, longitude, and CO variables.