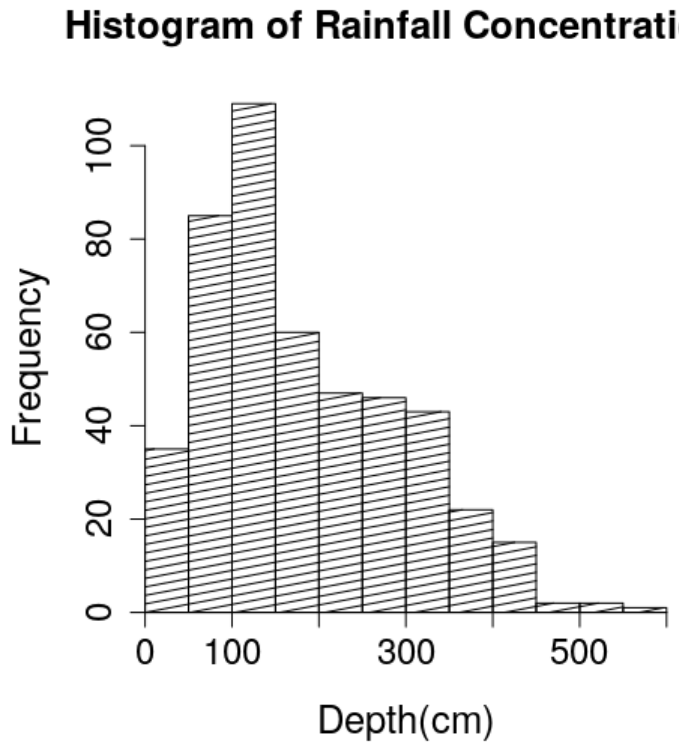
Stats 544 - Homework 1

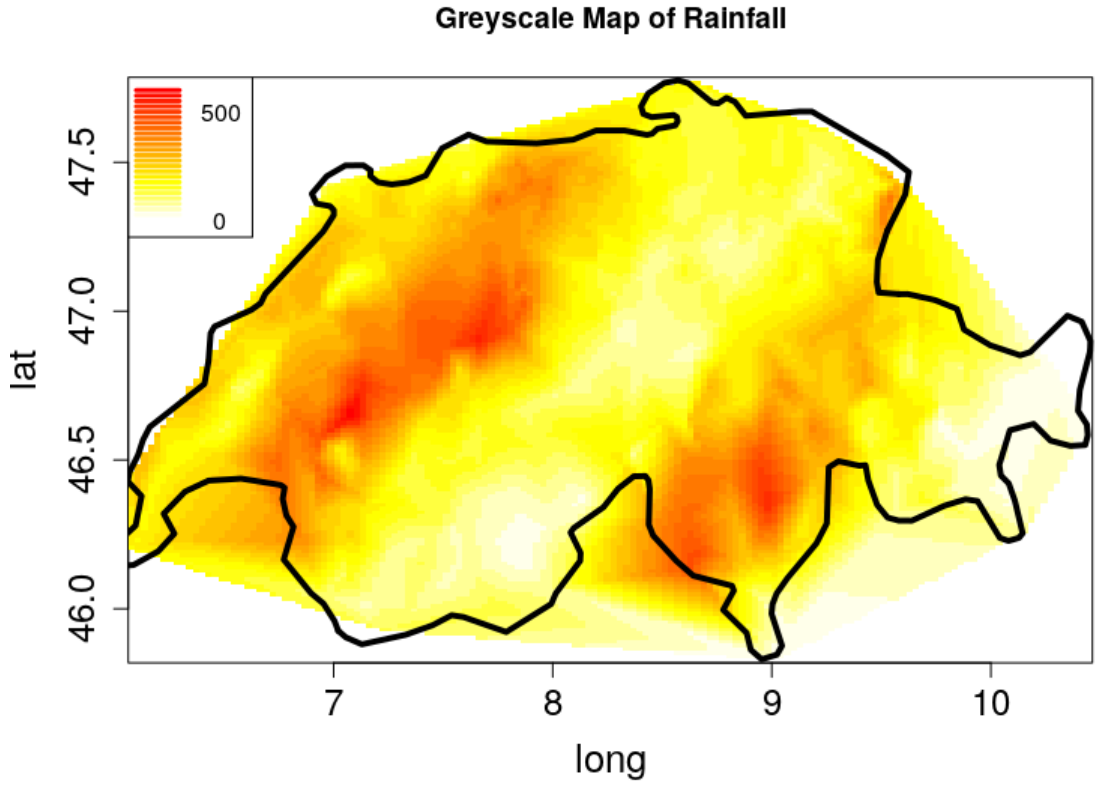
Franklyn Dunbar

Feb 3 2019

Switzerland Rainfall Data

467 daily rainfall measurements were made in Switzerland in May, 1986. This data is a sub-set of a continuous sample, and is uni-variate. Upon creating a histogram of the rainfall during a 24 hour window, we can see that the measurements are normally distributed around 150cm in depth. The calculated mean, standard deviation and range of the data are 184.24 cm, 112.26 cm, and (0,585) cm respectively. By interpolating the measurements and plotting the interpolating via heat map, we can analyze an estimate on what the continuous rainfall surface is. Though outlets strongly alter interpolation, the heat map provides a general idea of what the rainfall is over the total area. We can see in the heat map that there is a rainfall low up the middle of Switzerland trending to the north-east, with two concentrated blogs along the margins of the low that trend in the same direction. My guess is that this is a product of topography, given the distinct geometry. Another method to visualize the interpolation is with an indicator plot. This shows the percentage of the interpolated points that are above a certain threshold. This plots are sort of a binary heat map and show similar trends; a north-east trending low blob across the country and zones of higher rainfall along it's margins.


Histogram of Rainfall Concentrations
Frequency
0
20
40
60
80
100
0
100
300
500
Depth(cm)


Greyscale Map of Rainfall
500
0
lat
46.0
46.5
47.0
47.5
7
8
9
10
long

Rainfall % >58.6

Rainfall % >89.2

Rainfall % >114.8

Rainfall % >131.4

Rainfall % >152

Rainfall % >193.2

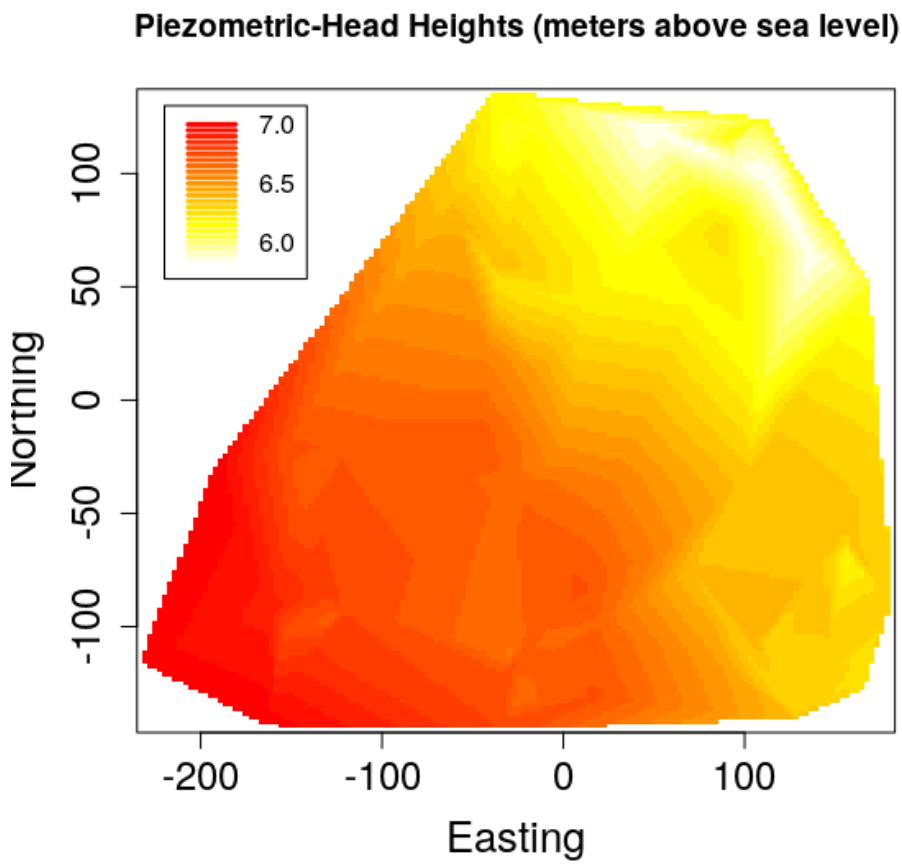
Rainfall % >234

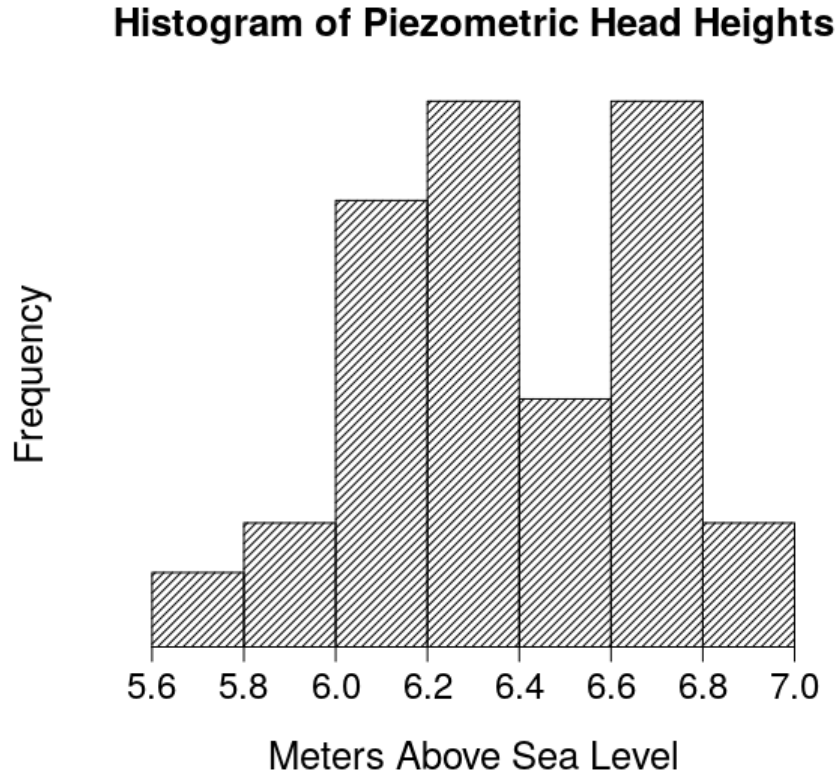
Rainfall % >288.8

Rainfall % >346.4

Wolfcamp Aquifer Data

The Wolfcamp Aquifer data is a sample of piezometric-head heights (meters above sea level) for an aquifer is West Texas. These 85 locations are a representative sub-set of a continuous random function and are univariate. The mean, standard deviation and variance of the data are 610 m, 186.33 m, and (312, 1088) m respectively. To decide if a transformation of the data was necessary, i used the so-called moving-window method to capture regional variances in order to see if outliers would significantly impact analysis. For a window size of 70x70 and an overlap of 10x10, the analysis returned a range of standard deviations from 25 to 123. Given this range of standard deviations, I transformed the data by taking the log in order to correct for the outliers. Given the relatively small range of measurements (They are all within 1 order of magnitude difference) it did not seem unreasonable to leave the data as is. Looking at a histogram plot of the heights, we can see that the data is somewhat normal (assuming a prettier bell curve with more bins), and is centered around 650 meters above sea level. As with the rainfall measurements from Switzerland, I interpolated the data to provide an estimate of what the random functions surface looks like. We can see in the heat map that the head height increases towards the bottom left, with a strong trend going diagonal across the interpolated space. In order to analyze the directional trends, I employed a correlogram to see how the respective correlation, covariance, and semi-variance changes along the x and y axes. The data is somewhat noisy, but in general we see a decreasing correlation along both axes with an "h" of 20, but slightly more correlation along the x axis.


Piezometric-Head Heights (meters above sea level)
7.0
6.5
6.0
Northing
100
50
0
-50
-100
-200
-100
0
100
Easting


Histogram of Piezometric Head Heights
Frequency
5.6
5.8
6.0
6.2
6.4
6.6
6.8
7.0
Meters Above Sea Level

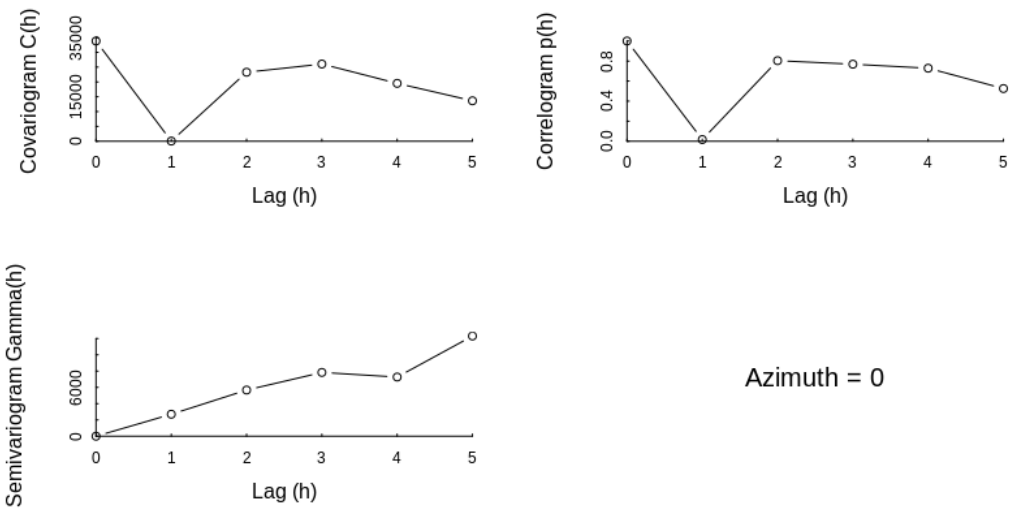


Figure 1: lag of 20 along y axis

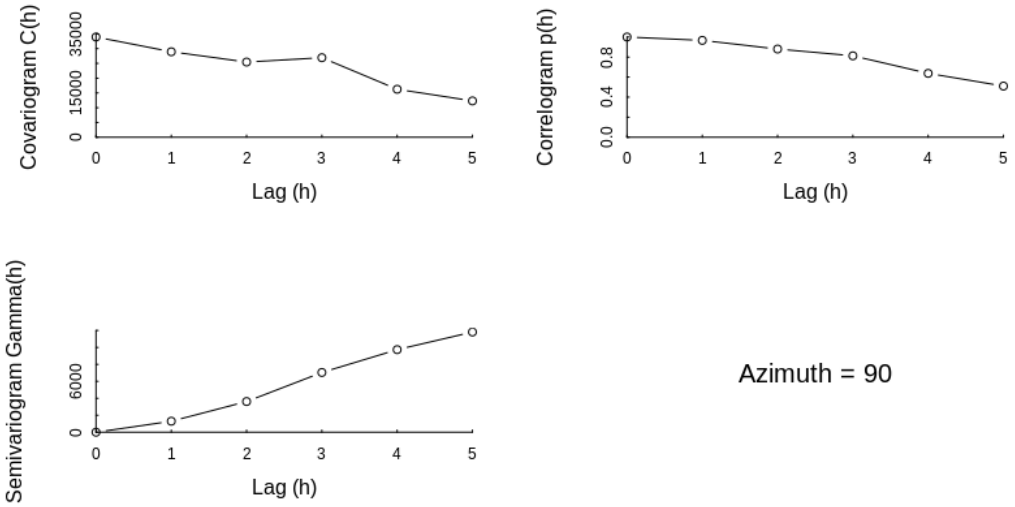
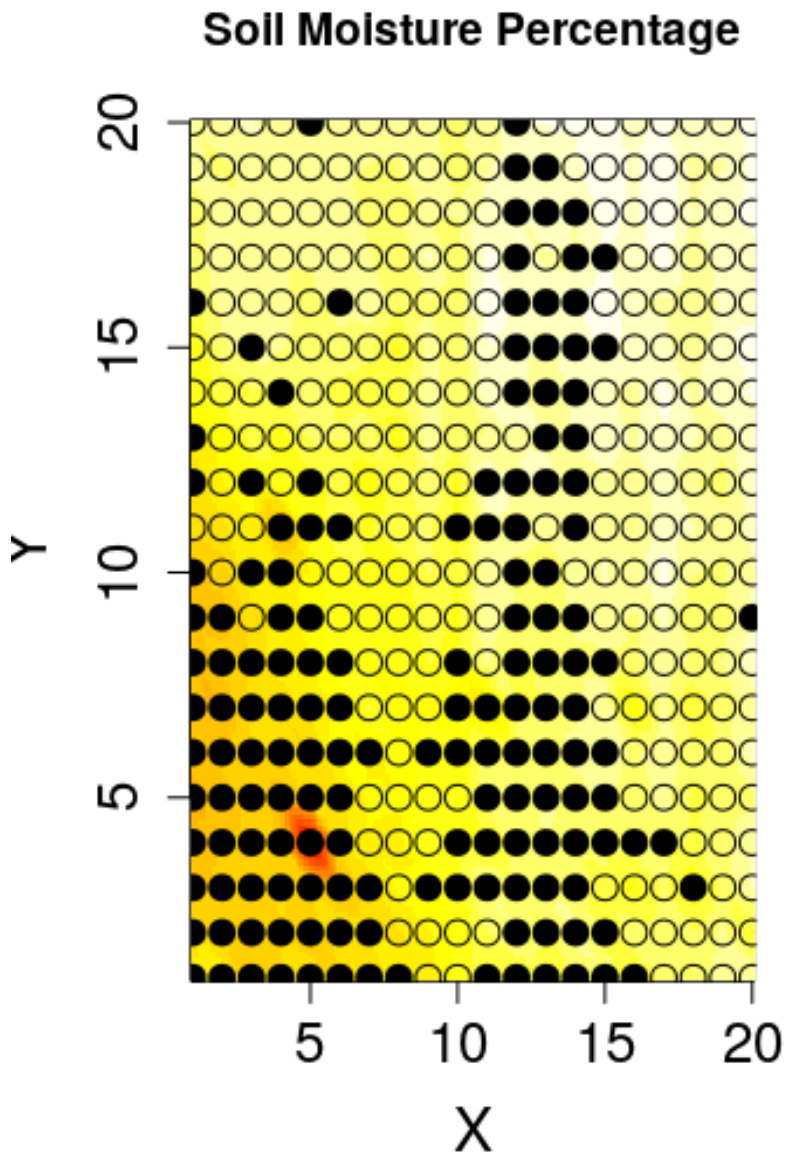


Figure 2: Lag of 20 along X axis

Phytophthora Data

The Phytophtora data is the recorded presence and absence of the disease Phytopthora among bell pepper plants and their corresponding soil water percentage. For my analysis I treated this data as bi-variate (diseased and water percentage), and both latticed data (disease instances)/continuous data (water percentage). The goal of this EDA was to see if there are any trends in disease instances and soil water percentages. To do this I interpolated the soil water percentages, and then plotted disease instances over it.

The mean and standard deviation in moisture percentages is 8.8%, and 2.38% respectively with a range of (5.23, 26)%. The percentage of diseased instances is approximately 62%. To explore correlation between disease instances and soil moisture percentages, a correlogram was created. We can see in the plots below that there is not much correlation (around .4 along both axes) between disease instances and soil water percentage along either exes.



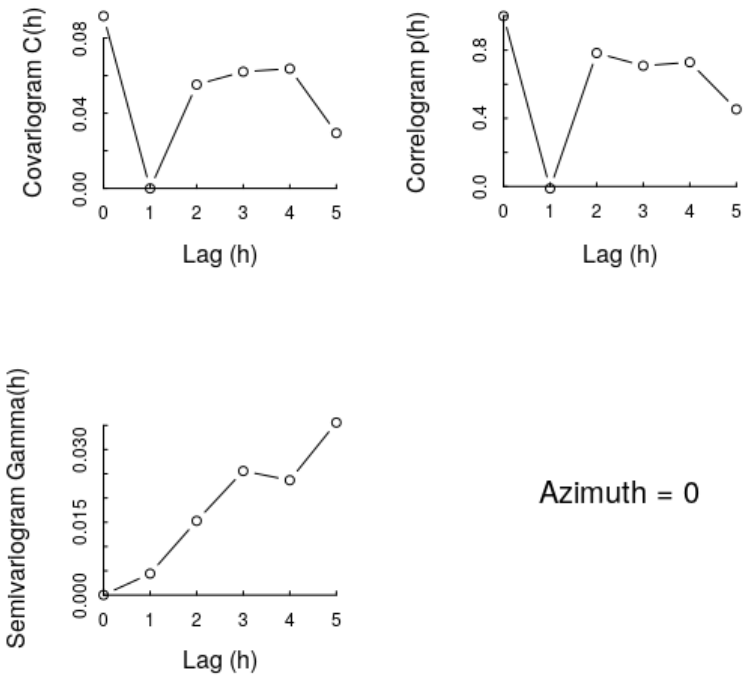


Figure 3: lag of 1 along y axis

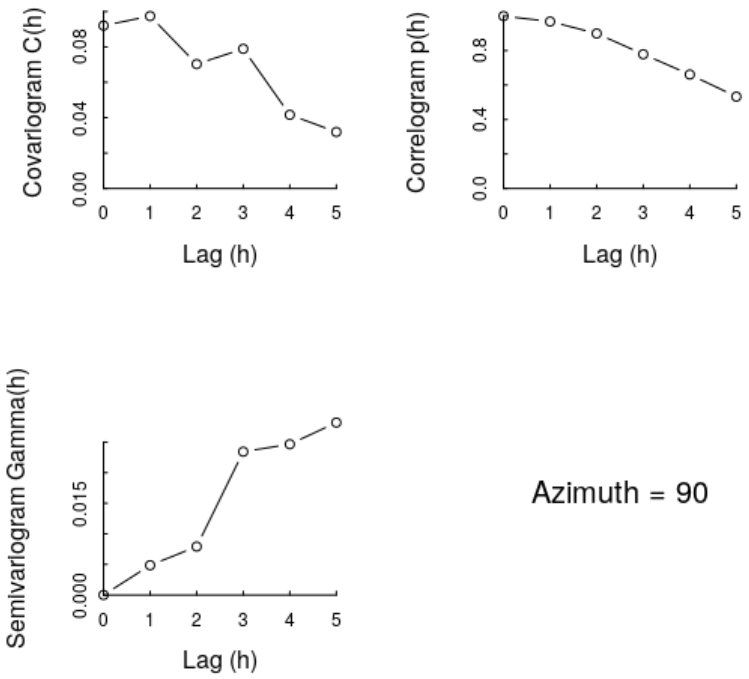
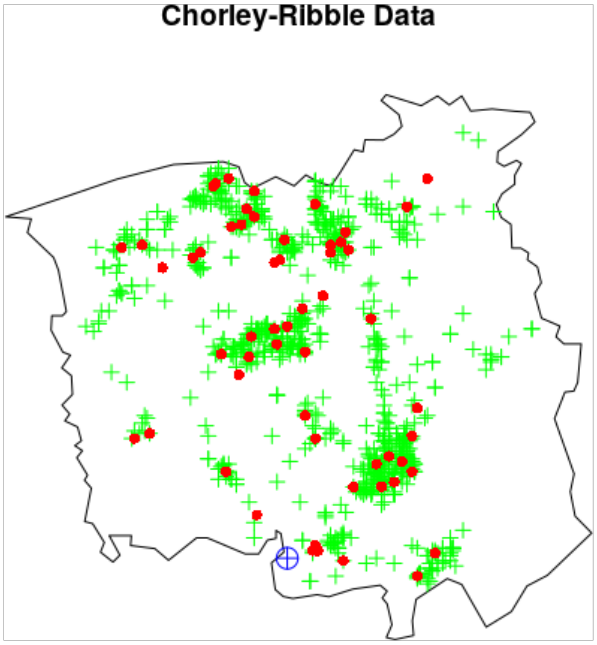


Figure 4: Lag of 1 along X axis

Chorley Lung/Larynx Cancer Data

This data is clearly an instance of a point process, given that we have no idea what the frequency is of the cancer instances relative to the population. My initial EDA was to plot the cancer instances relative to the location of the incinerator to see if one could eyeball any obvious clustering near it. There were three goals in my analysis, to determine if instances of larynx cancer were clustered, to determine if instances of lung cancer were clustered, and see if there was correlation between the distance from cancer instances and the incinerator were correlated (i.e. does the incinerator cause cancer?).

The percentage of lung cancer instances from the total instances of cancer is approximately 94%, which shows an obvious bias towards lung cancer. Quadrat tests on locations of both instances yielded a p-value of 0.0015 for larynx cancer, and 5e-4 for lung cancer. While these are both indicative of clustering, it could be argued that instances of lung cancer are more clustered than instances of larynx. To determine spatial correlation between all cancer instances and the incinerator, a correlogram was created to measure the relationship between cancer instances and radial distance from the incinerator. This was done along both axes for both larynx and lung cancer. The resulting plots show that there the values for both cases correlate along the x axis, but not necessarily along the y axis.



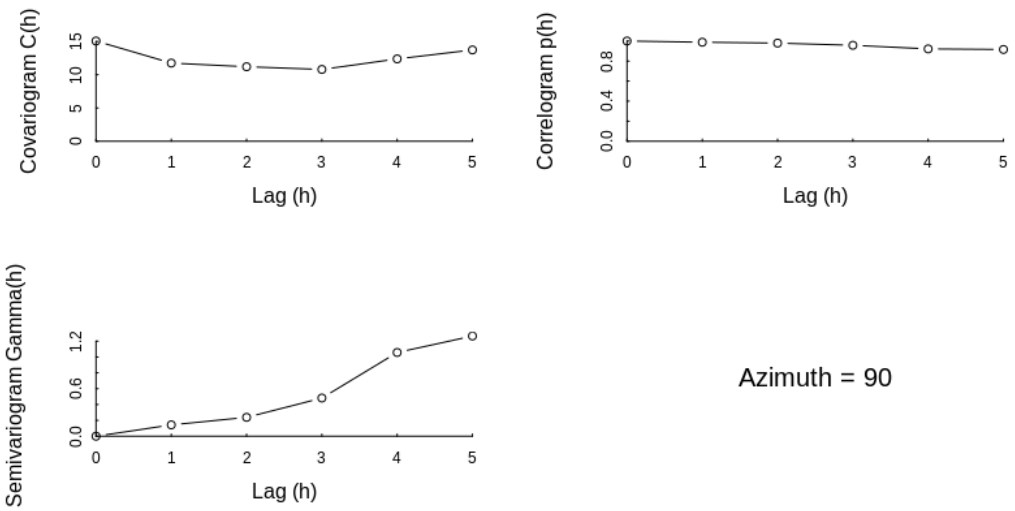


Figure 5: lag of 1 along X axis

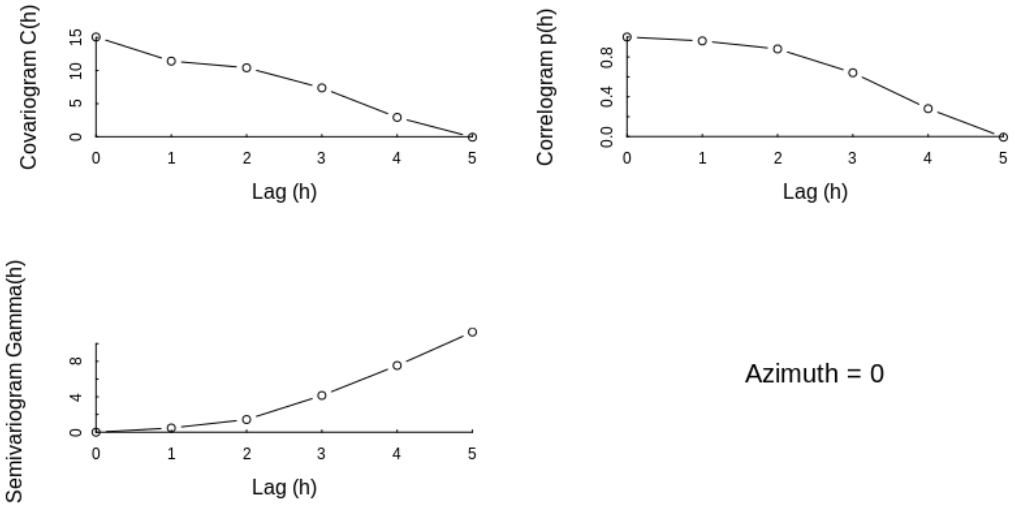


Figure 6: Lag of 1 along Y axis

Figure 7: Correlograms for Lung Cancer

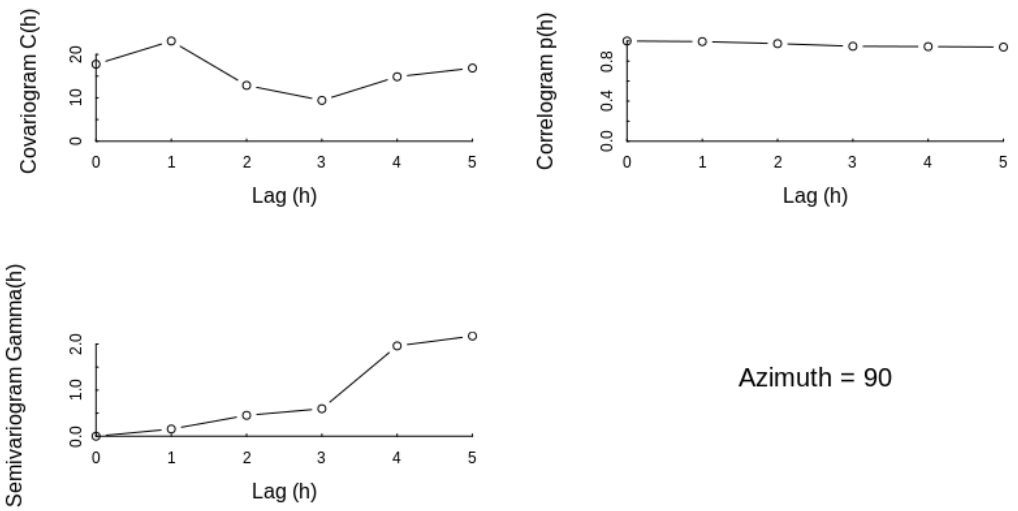


Figure 8: lag of 1 along X axis

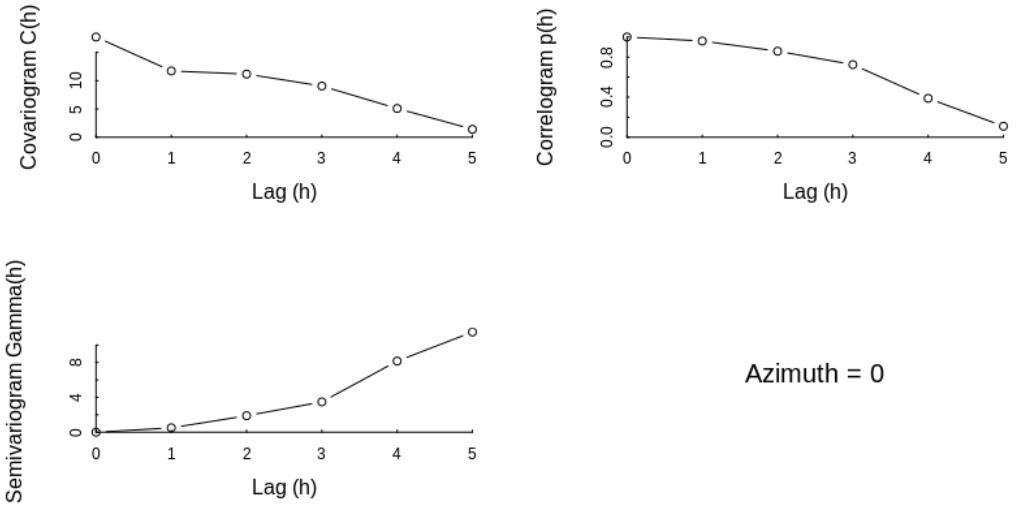


Figure 9: Lag of 1 along Y axis

Figure 10: Correlograms for Larynx Cancer