Geo Statistics Final Project

Aidan O'Sullivan

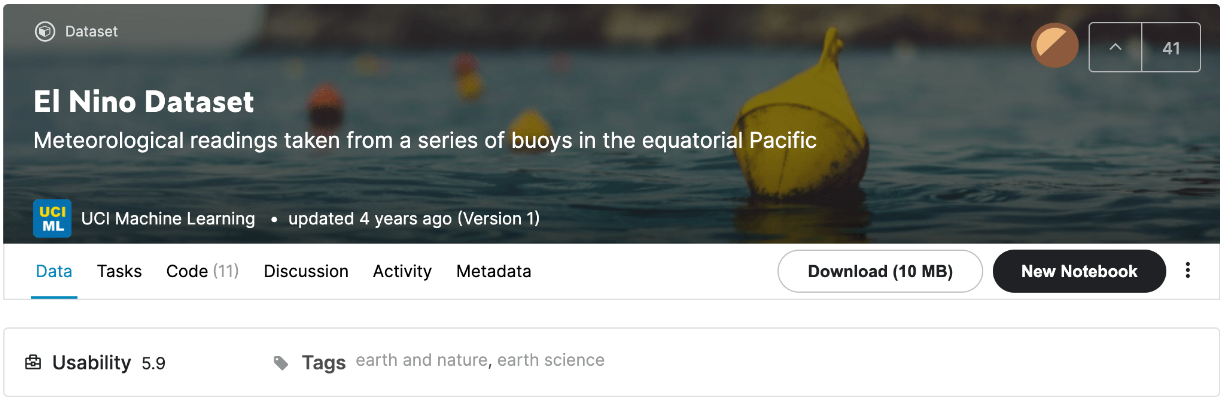
3/10/2021

Introduction

For my final project, I chose to analyze buoy data around Malden Island, just south of the Equator in the central Pacific. These buoys, which help researchers track trends used for El Nino and La Nina prediction, monitor features such as Zonal Wind, Meridional Winds, Humidity, Air Temperature, Sea Surface Temperatur, and of course longitude and latitude.

After initial analysis, I decided to predict relative air humidity, as this was the most normally distributed variable available that also has a noted scientific impact as a greenhouse gas and foreteller of precipitation. Relative Humidity is measured as the current point pressure in the air divided by the pressure point at which water vapor saturates the air, then multiplied by 100 to create a percentage. The higher the percentage, the more moisture resides in the air.

Data Source



Dataset

El Nino Dataset

Meteorological readings taken from a series of buoys in the equatorial Pacific

UCI Machine Learning • updated 4 years ago (Version 1)

Usability 5.9

Tags earth and nature, earth science

Figure 1: Kaggle El Nino Data

I collected my data from Kaggle's explore page, though its ultimate source is the University of California at Irvine Machine Learning Data Archive This particular dataset tracked tens of thousands of buoys across the equatorial Pacific from the years 1992 through 1997.

Location

Due to the initial size of the dataset - 1780,000 observation from tens of thousands of buoys - I narrowed my point of interest to Malden Island, and an uninhabited "recursive" island located along popular shipping routes. Within 10 nautical miles of the island, there are 55 unique buoy observations between the years 1992 and 1997. Because the buoys move ever so slightly from year to year and the the data between these particular years was so similiar, I decided not to perform time series analysis and just view the five year collection period as a single, long-term period of data collection.

With the narrowed down dataset, I then set to work checking the various distributions of target variables, the correlation between then, as well as outliers. In this report, I will focus upon Humidity, as that became the target variable of choice.

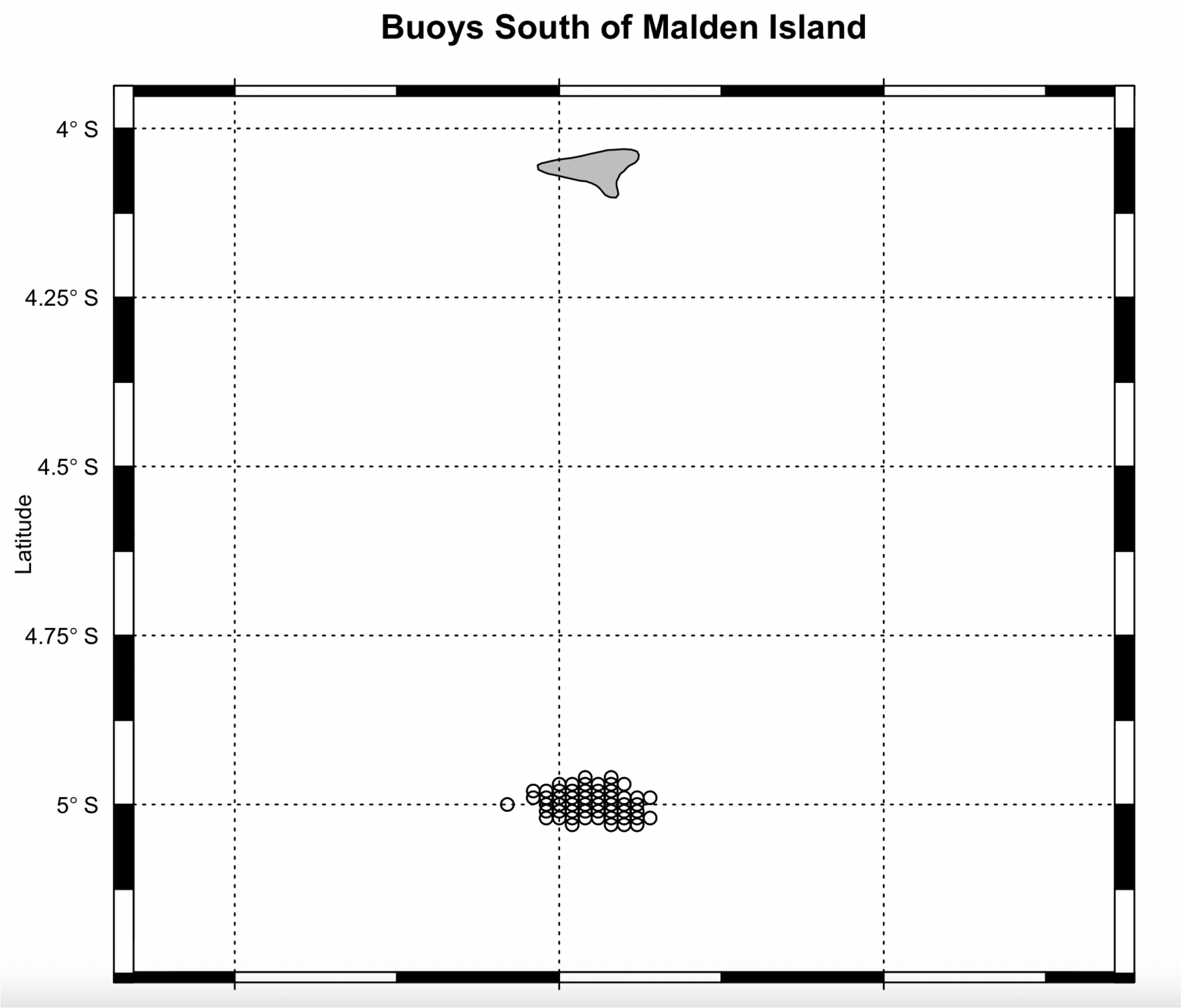
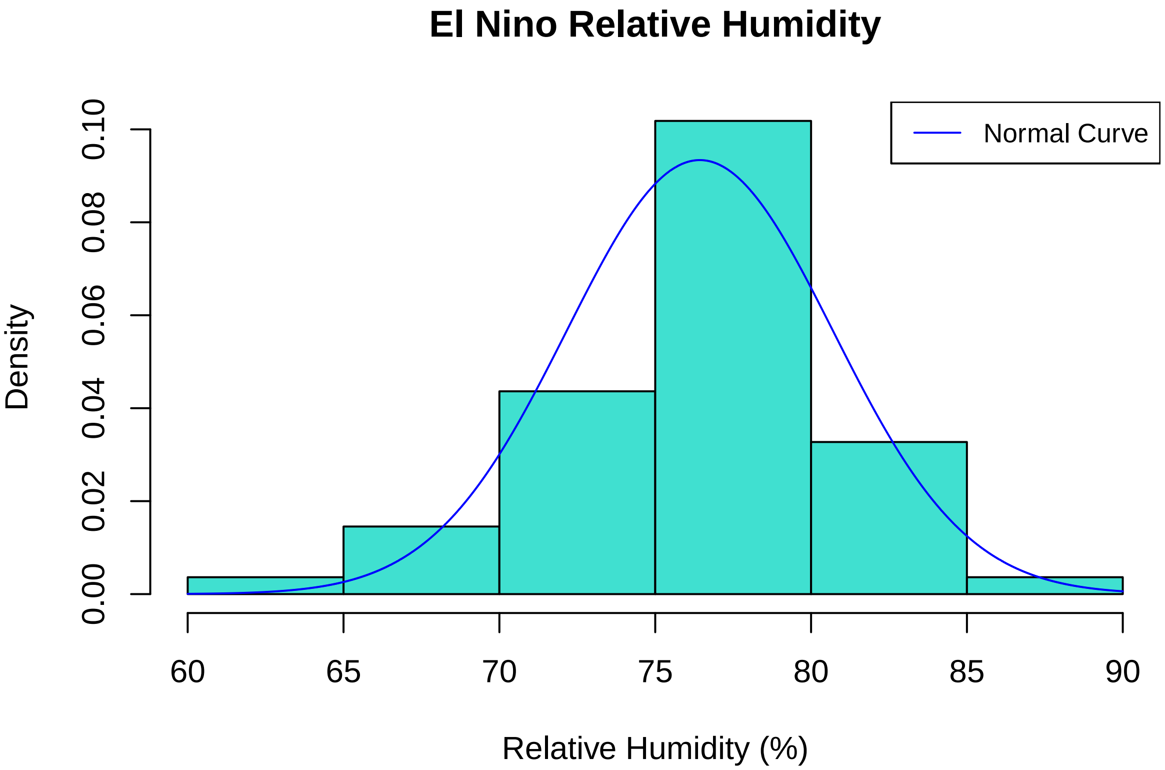
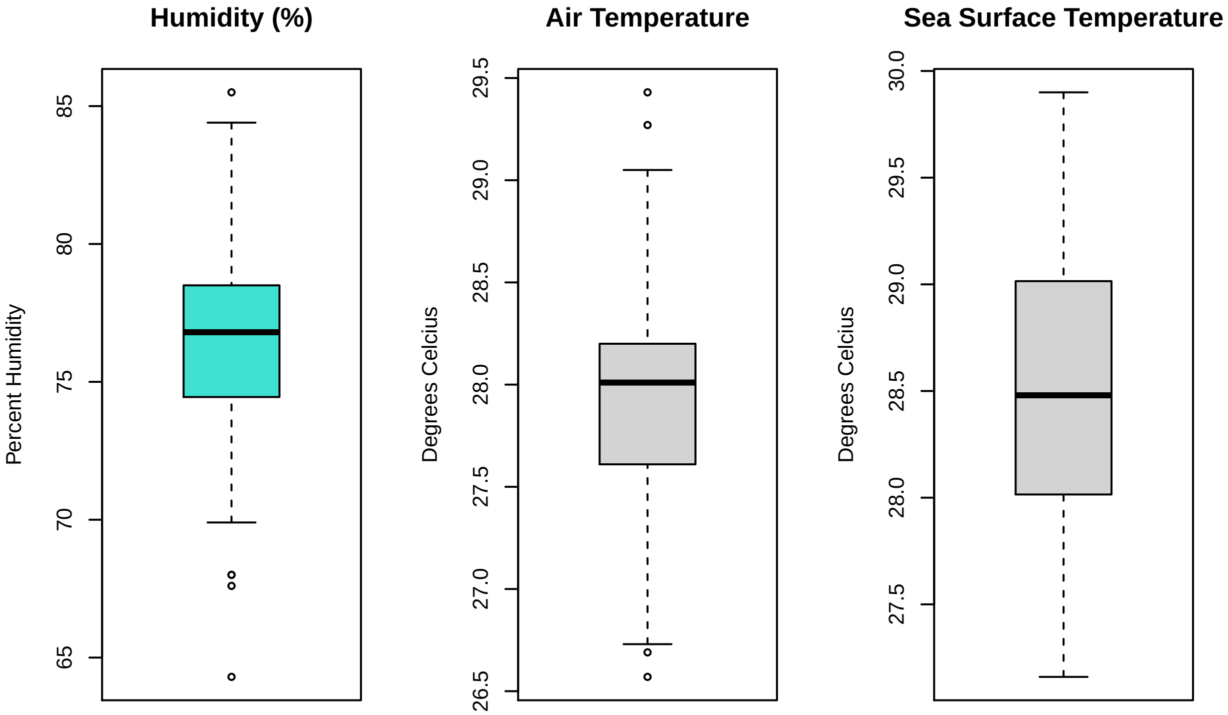


Figure 2: Malden Island Buoys

Data Exploration



As we can see from the above history, Relative Humidity across the 55 loctions is extremeley normall distributed, even given the small sample size.



Next, by looking at the boxplots of Humidity, Air Temperature, and Sea Surface Temperature, we can see there are very few outliers. In fact, regarding Humidity, the only notable outliers are those observations with low humidity. Fortunately, as we can see from the point plot below, these observations of low humidity are clumped together such that they add interesting variation to the data. In other words, they are points of high leverage.

Plot of Buoys by Humidity

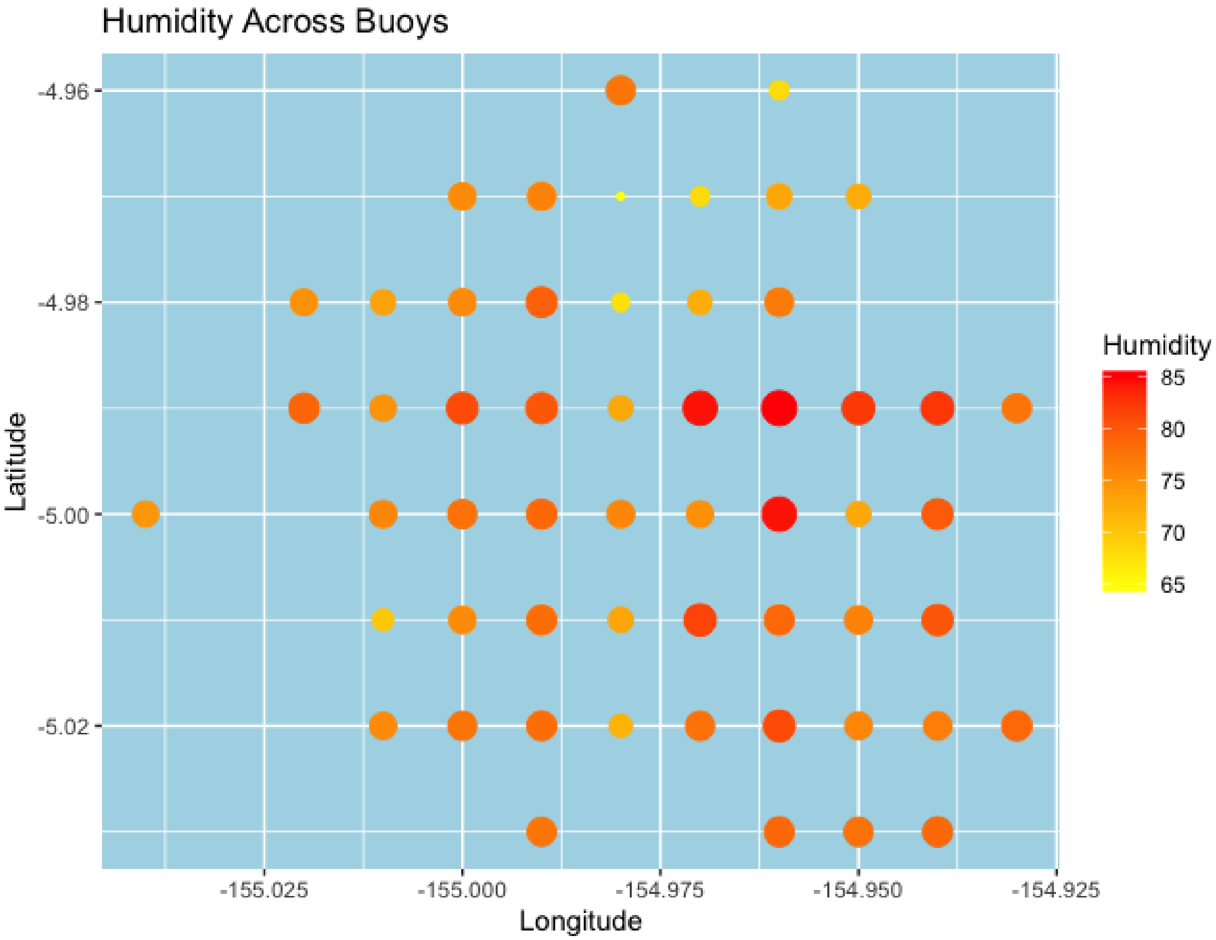


Figure 3: Humidity Across Buoys south of Malden Island

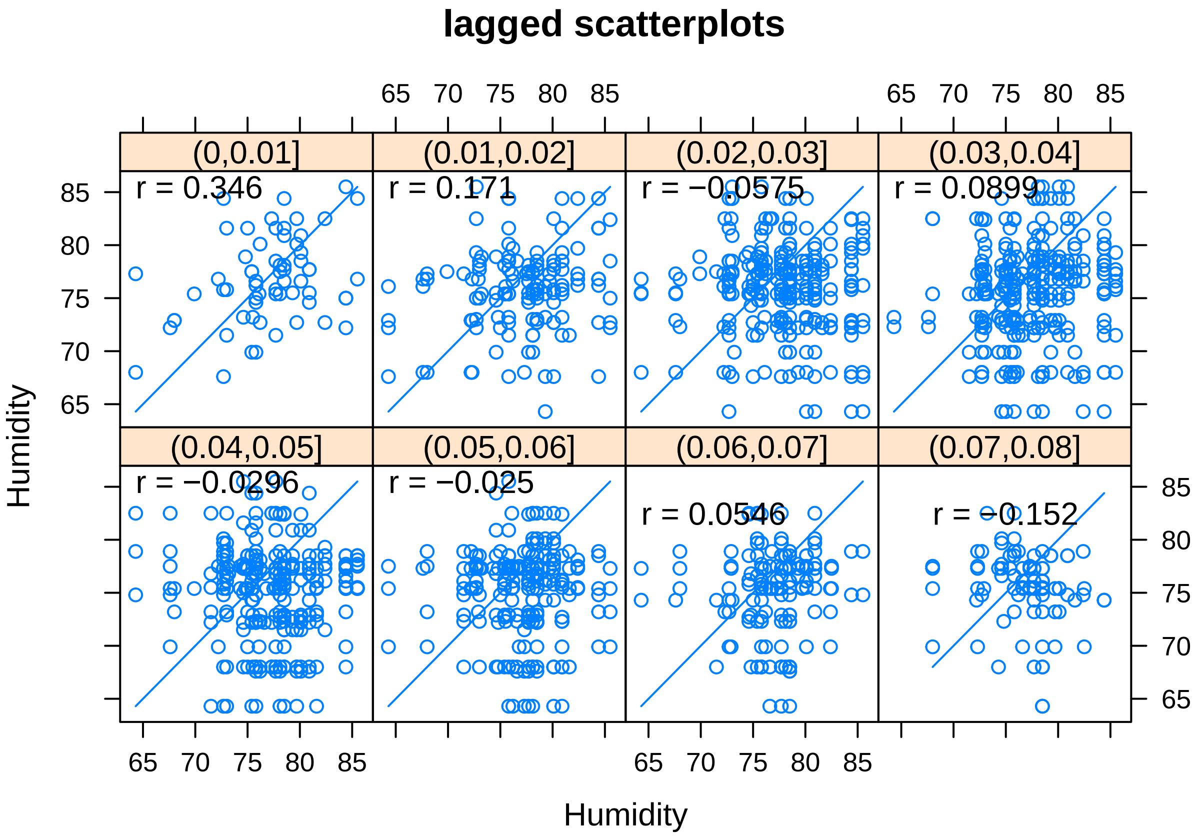
Correlation

General Correlation

```
##                  Zonal.Winds Meridional.Winds Humidity Air.Temp
## Zonal.Winds                1
## Meridional.Winds        0.13                1
## Humidity                0.16            -0.03        1
## Air.Temp               -0.08             0.21    -0.26        1
## Sea.Surface.Temp        0.38             0.18    -0.02     0.74
##                  Sea.Surface.Temp
## Zonal.Winds
## Meridional.Winds
## Humidity
## Air.Temp
## Sea.Surface.Temp                1
```

The only notable correlation found in the dataset is between Air Temperature and Sea Surface Temperature, which makes a lot of sense. Zonal winds and Sea Surface Temperature have the next highest correlation, though not even breaking the 0.5 mark. Lastly, Humidity appears negatively correlated with Air Temperature, which also makes sense given how Humidity is calculated. This does not bode well for future co-kriging.

Spatial Correlation



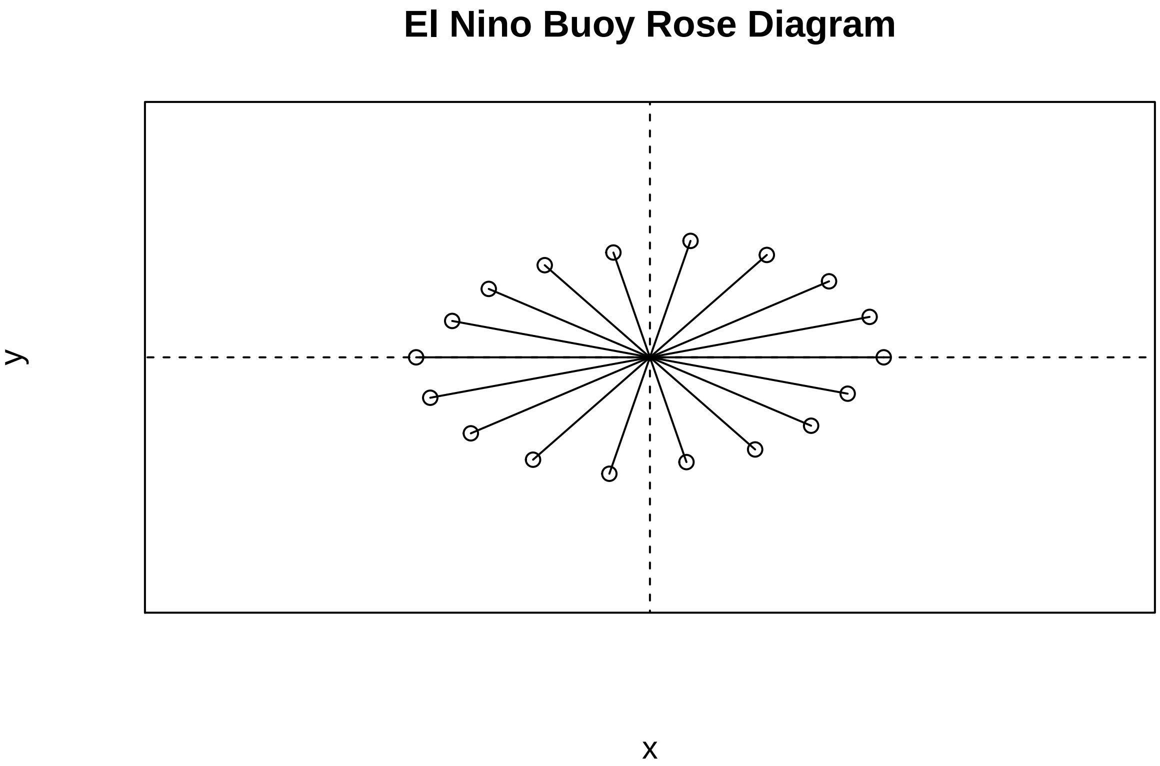
Lastly, we turn to the h-scatterplots to quantify spatial correlation at various distances, measured generally in units. In this case, the distances are extremely minimal since the buoys are so close together; additionally, the overall correlation is small as well. Even at the closest distance, correlation is only about 0.346 and falls to nearly 0 as the distances increase.

Methodology

Variograms

Now that the data has been thoroughly explored, the next step of the geostatistical analysis is to understand the anisotropy of the buoy data, determine trends in the data, and finally plot and fit variograms

Rose Diagram - Checking for Anisotropy



Given the nearly circular shape of the Rose diagram, we can conclude that the data is not direction-dependent, but rather direction-independent. In other words, the data is more or less isotropic, not anisotropic. No adjustments must be made to the data.

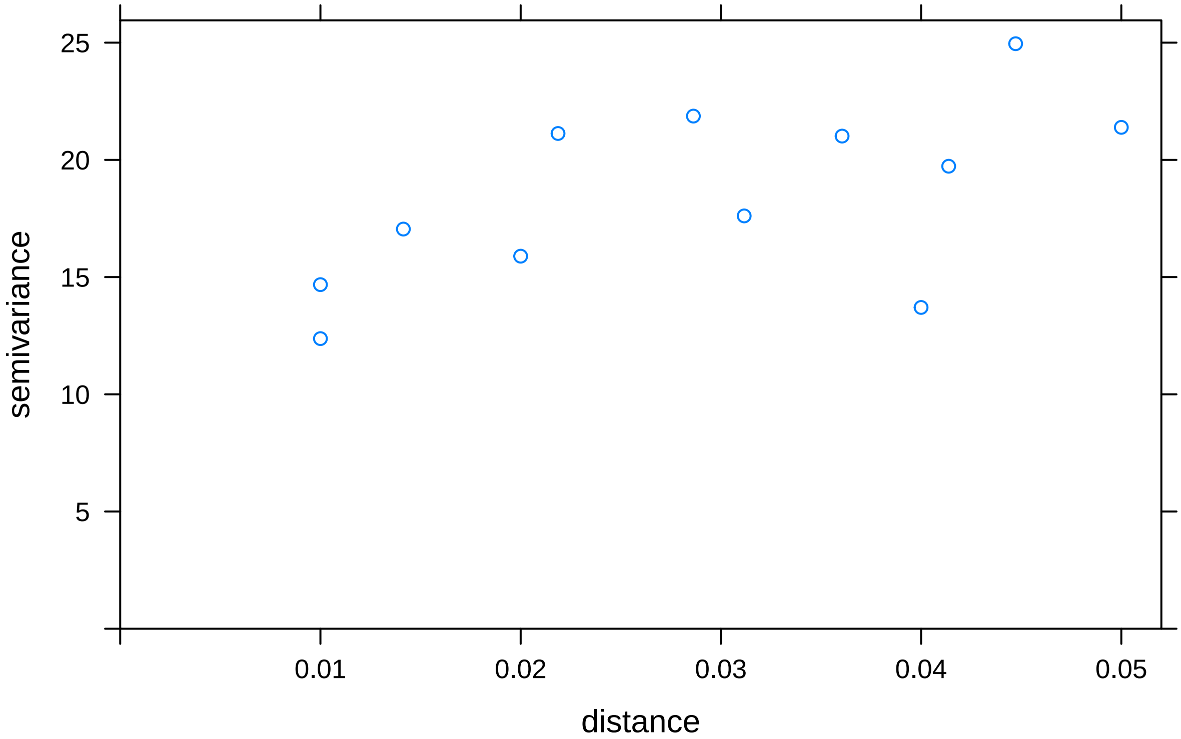
Fit Model Variogram

The next step is to fit a model variogram. We will begin by using the normal estimator, and choose between a spherical model, exponential model, and linear model.

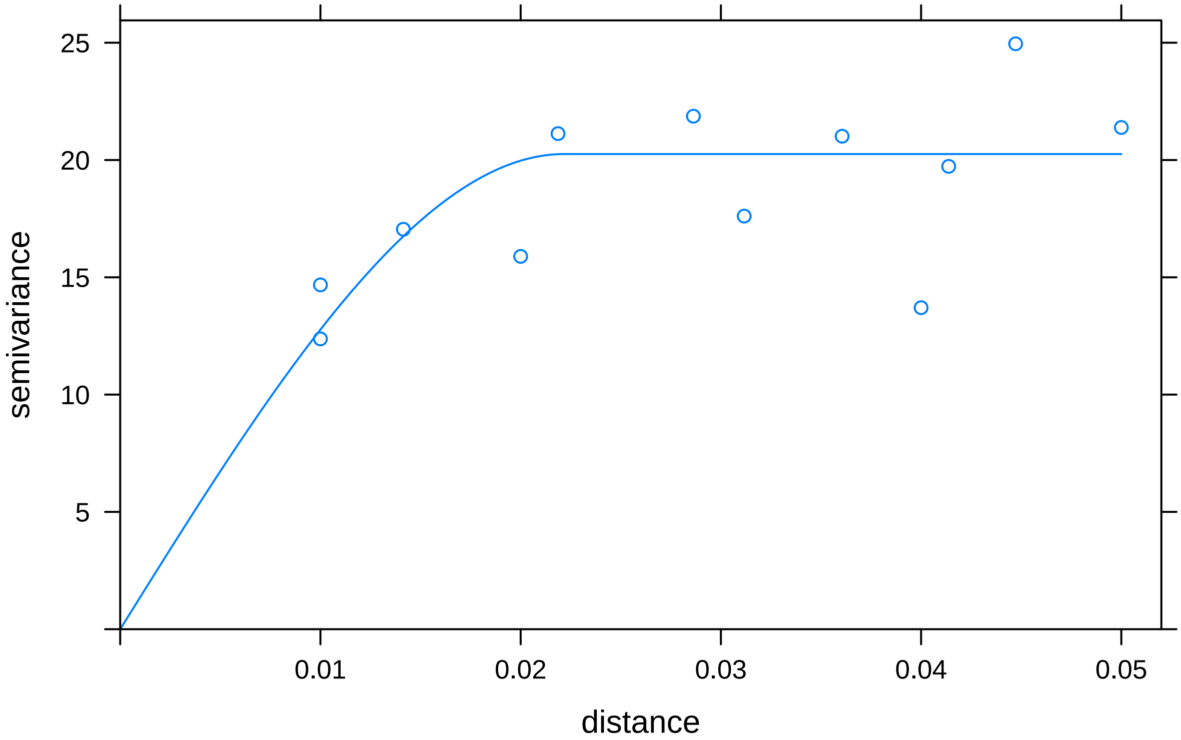
```
##   model    psill     range
## 1   Sph 20.25474 0.0221758
```

Based on the several model options available, the best model variogram is the Spherical model with a sill of 20.25 and a range of 0.022 units.

Sample Variogram, Normal Estimator



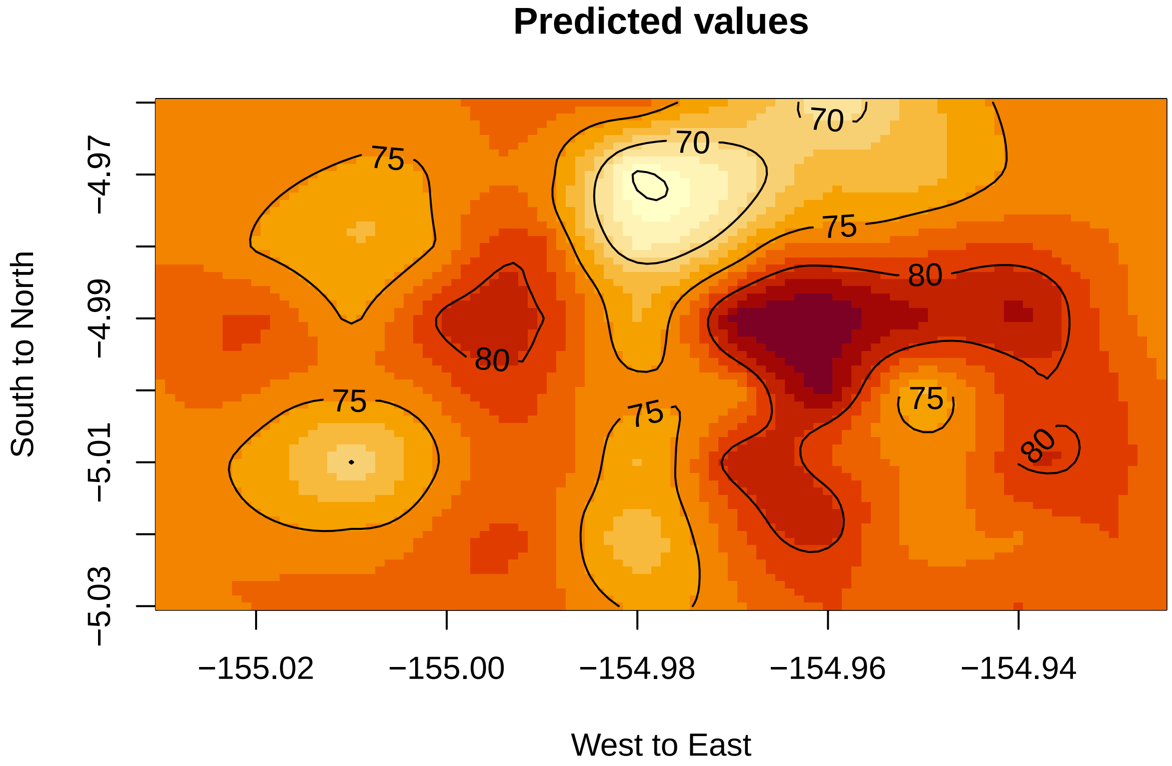
Fitted Sample Variogram, Normal Estimator



Kriging

Next, I will perform basic kriging along with co-kriging before using cross validation to choose between models.

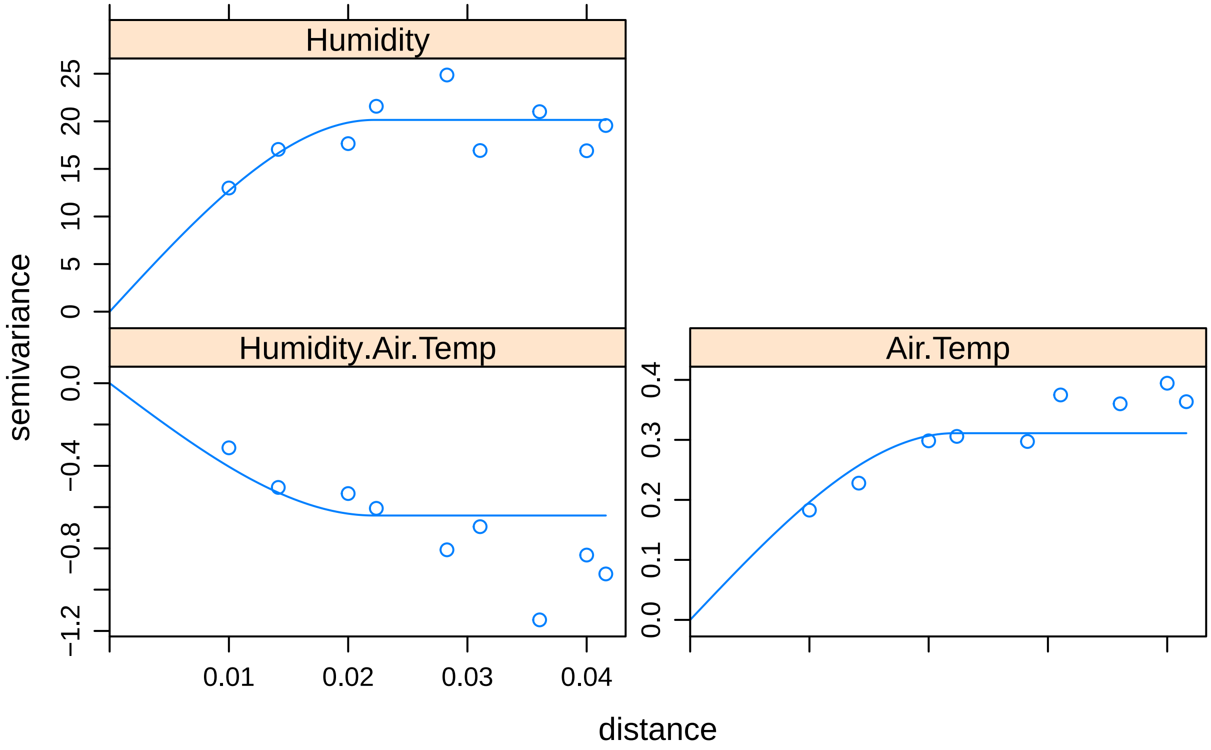
I will perform oridinary kriging using the best fit spherical model found previously.

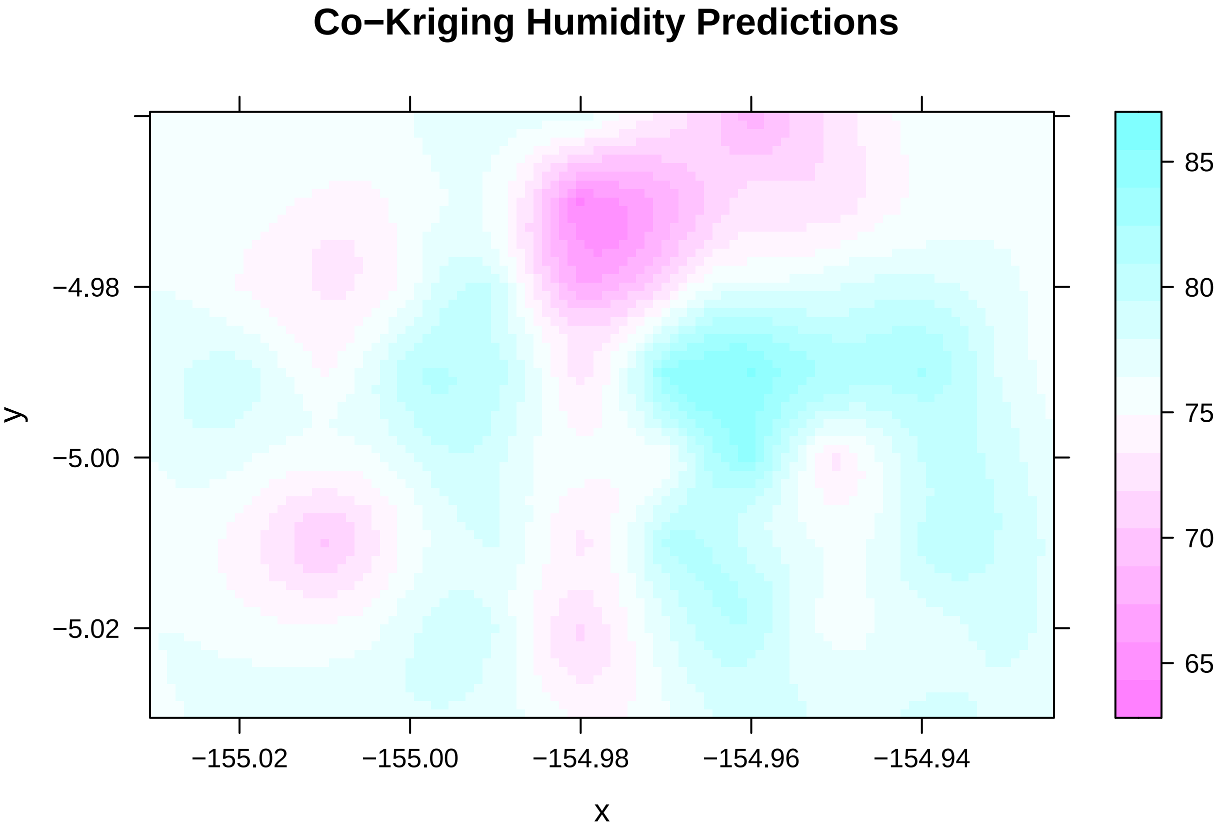


From this heat map, we can see that the buoys around -4.99 South and -154.96 West predict the highest humidity in the air. Just to the northwest of this spot, the buoys predict the lowest relative humidity in the microclimate. There is a possible corridor of lower humidity down the center of spine of the buoys.

Co-Kriging

In effort increase prediction accuracy but reducing prediction variance, we will now consider co-kriging, a system of kriging that harnesses correlation between a target variable such as Humidity and other co-target variables. In this case, I selected Air Temperature as the only co-target. Air Temperature is negatively correlated with Humidity by a correlation value of -0.26. This is a weak connection, but may still serve to decrease the PRESS score of the model.





The co-kriging also exactly replicated the odinary kriging model.

Cross Validation

Cross validation is a useful process to compare the accuracy of various models on a dataset while also reducing the risk of overfitting.

For the cross validation exercise, I will compare several different models:

1. Ordinary Kriging with Normal Estimator
2. Ordinary Kriging with Robust Estimator
3. Universal Kriging with Normal Estimator
4. Universal Kriging with Robust Estimator

Additionally, I will compare the winner of these models against a Co-Kriging Model.

```
##                               [,1]
## PRESS Normal, Ordinary  12.71978
## PRESS Normal, Universal 12.86983
## PRESS Robust, Oridinary 12.71978
## PRESS Robust, Universal 12.71978
```

After the first round of Cross-Validation, we can that there is very little difference in PRESS score between the many model. In fact, all the models have the exact same PRESS score except the Universal Kriging model with the normal estimator. Therefore, we can assume that the outliers in the dataset do not decrease the accuracy of the model, and that in future test we should use the best, simplest model availbe - the Ordinary Kriging model with the normal estimator.

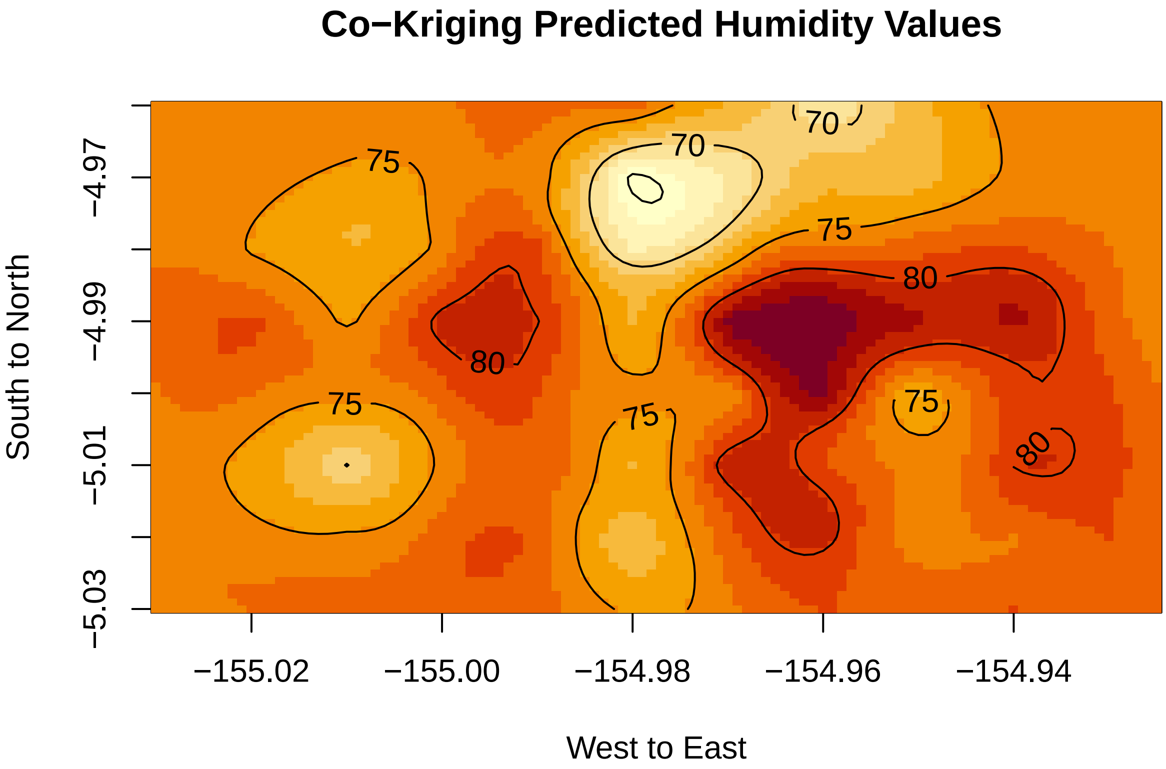
```
##                              [,1]
## PRESS Co-Kriging       12.25933
## PRESS Normal, Ordinary 12.71978
```

Next, we will use Cross Validation to compare the winner of the previous round - Ordinary Kriging with normal estimator - with the co-kriging model. Here, we have a clear winner. The Co-kriging model successfully used what little correlation was available and the PRESS value fell accordingly

Results

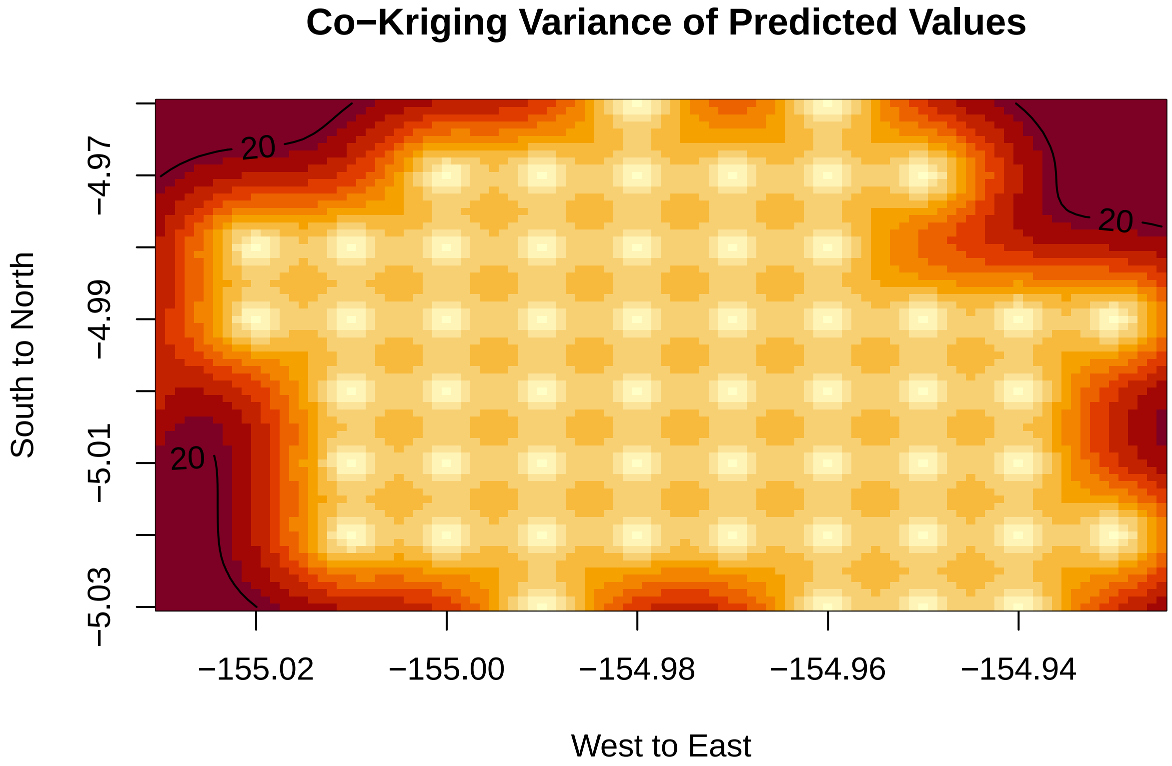
Now that we know that co-kriging is the best model, we use it to make final predictions.

Final Predictions



The final prediction model is almost indistinguishable from earlier models, even though mathetically speaking it is more accurate.

Final Prediction Variances



The variance of the predictions increases drastically as the predicted points move away from the observed buoys. This is why the grid of prediction has been reduced to area directly around the buoys and excludes even the sea around Malden Island.

Conclusion

This report concludes that a co-kriging model of Humidity and Air Temperature with a normal estimator is the best model to predict Humidity in the sea region south of Malden Island. Also, the spherical model is the best variogram model fit for the data.

Further research should be performed to extrapolate these findings to a larger radius of interest. This dataset of buoys collecting data between 1992 and 1997 across the entire equitorial Pacific Ocean only inlcudes dense collections of observed points separated by hundreds of miles of open ocean, a horrible combination for spatial prediction. Still, this research may be useful in determining interesting patterns of correlation between Humidity and sea life, or possibly tip off workers to malfunctioning buoys.